

R E P O R T R E S U M E S

ED 003 365

24

THE ANALYSIS OF PUBLIC CONTROVERSY, A STUDY IN CITIZENSHIP
EDUCATION.

BY- OLIVER, DONALD W. AND OTHERS

HARVARD UNIV., CAMBRIDGE, MASS., LAB. RES. IN INSTR.

REPORT NUMBER CRP-551-VOL-2

PUB DATE

62

HARVARD UNIV., CAMBRIDGE, MASS., GRAD. SCH. OF EDUC.

EDRS PRICE MF-\$0.45 HC-\$11.00 275P.

DESCRIPTORS- SOCIAL STUDIES, SOCIAL VALUES, *SOCIAL PROBLEMS,
*CITIZENSHIP, PUBLIC SCHOOLS, TEACHING PROGRAMS, PROGRAM
EVALUATION, *CURRICULUM DEVELOPMENT, *POLITICAL ISSUES,
*SOCIAL SCIENCES, CAMBRIDGE, MASSACHUSETTS

A DISCUSSION WAS CONTINUED IN THIS REPORT ON
IMPLEMENTING AN APPROACH AND ON DEVELOPING AND EVALUATING A
TRIAL PROGRAM FOR USING ISSUES OF PUBLIC CONFLICT AND
POLITICAL CONTROVERSY AS CONTENT ELEMENTS IN PUBLIC SCHOOL
SOCIAL SCIENCE PROGRAMS. CONCLUSIONS WERE PROVIDED FOR THE
VARIOUS TOPICS DISCUSSED. THE INFORMATION WAS PRESENTED AS
VOLUME TWO OF A TWO-PART REPORT. (REFER TO ACCESSION NUMBER
ED 003 364, VOLUME ONE) (JH)

Cooperative Research Project, 8145
(Contract #SAE-551)

THE ANALYSIS OF PUBLIC CONTROVERSY
A STUDY IN CITIZENSHIP EDUCATION

by

Donald W. Oliver

and

James P. Shaver

U. S. DEPARTMENT OF HEALTH, EDUCATION AND WELFARE
Office of Education

This document has been reproduced exactly as received from the
person or organization originating it. Points of view or opinions
stated do not necessarily represent official Office of Education
position or policy.

Harold Berlak

With the assistance of:

and

Ernest Van Seasholes

The research reported herein was performed
pursuant to a contract with the United States Office of Education,
Department of Health, Education, and Welfare

Harvard Graduate School of Education
Harvard University
1962

ED 003365

CHAPTER TEN

EVALUATING COMPETENCE IN THE ANALYSIS OF POLITICAL CONTROVERSY

One of the most difficult problems faced by anyone who wants to deal systematically with the problem of teaching students to analyze political controversy critically is the selection and development of instruments for evaluating competence in this type of analysis. This is especially difficult if one's main interest is not simply in assessing the student's ability to memorize or comprehend definitions of analytic concepts, but rather if we try to measure the student's ability to use the concepts to evaluate new and complex situations. The Project was centrally interested in discovering or developing measures of such analytic skills or abilities in order to evaluate the results of teaching an experimental curriculum and variation in teaching style. Our search through the literature was, quite frankly, disappointing. Our attempt to develop new measures, moreover, met with some rather awesome obstacles. Both of these efforts are reported in this chapter.

Some Major Measurement Efforts to Assess the Ability to Analyze Political and Social Issues

In searching for tests which were appropriate for measuring competence in political analysis we found that most fall under the rubric "critical thinking." We shall first survey some of the major efforts to assess "critical thinking." This, we hope, will allow the reader to see more clearly the general problems inherent in this type of measurement. After this review we shall describe some of our own efforts to meet these problems. We should note that although we shall present a discussion of existing measures of critical thinking, our purpose is not to relate them to any particular experiment at this point. Some of the measures discussed here will be discussed again in Chapter 11, where our concern will be with the way the particular measure was actually used in an experimental investigation.

The Watson-Glaser Critical Thinking Appraisal

One of the early, and significant, experiments investigating the teaching of "critical thinking" was conducted by E. M. Glaser.¹ To evaluate the results of his research, he employed a set of measures called the Watson-Glaser Tests of Critical Thinking (now entitled, the Watson-Glaser Critical Thinking Appraisal) which is still in wide use. The rationale behind the Critical Thinking Appraisal is explicated by Glaser when he discusses his broad goals for measuring the skills and attitudes necessary for good citizenship.

He is interested in developing a good citizen who: (1) understands our political and economic system, (2) can think critically about issues where there is disagreement, and (3) is committed to democratic values. Thus, the

¹Edward M. Glaser, An Experiment in the Development of Critical Thinking (New York: Teachers College, Columbia University, 1941).

1951 manual of the Watson-Glaser critical thinking appraisal explains one of the many possible applications of the instrument.

Perhaps the most important broad uses....are as an evaluation instrument and as a teaching tool to help students and trainees develop reliable techniques for logical reasoning that will guide them in their daily life situations. Ability to think critically is tied closely to the citizenship attitudes and skills needed to preserve and strengthen our democratic state.²

The measure. The 1942 version of the Watson-Glaser test is made up of six sections.

Tests A1 and A2 are an "agree-disagree type" attitude scale. It is intended to measure consistency of a person's belief on social issues. For example, does the student consistently favor government intervention on behalf of the less fortunate. In addition to a "consistency score" the test yields a "P" or progressive score. The progressivist position is defined as "supporting political and economic democracy, civil liberties, the welfare of all the people and equality of opportunity, while the "other" view is described as "reactionary, suppressive, and intolerant." A third or "U" score is based on the number of times a student checks "uncertain." A high "U" score indicates "ignorance or pedantry or resistance to the test," or ".... justifiable feeling that the evidence does not warrant a definite opinion."³ The 1952 edition omits this test.

Test B is intended to measure logical reasoning in both non-controversial and controversial areas. The items are of two types: a short syllogistic form, and a slightly longer paragraph form with a more elaborate presentation of evidence and situation. The situation is followed by three or four conclusions, and the student is to decide whether each conclusion "follows" or "does not follow." The 1952 version calls this test "Deduction" (Test 3), and omits the longer paragraph items.

Test C, Inference, is designed to measure the ability of the student to judge the truth, falsity and relevance of an inference drawn from a statement of fact. A statement is presented which is assumed to be true, a series of statements follows which are inferences or conclusions. The student must judge whether each of the inferences or conclusions is true, false, probably true, probably false, or whether there is insufficient data to make a judgment. The 1952 version of the inference test is essentially the same as the 1941 test. It contains situations dealing with controversial social issues (labor unions, segregation, freedom of press). The directions for taking this part of the test occupy one full page in which the subjects are, in effect, taught how to make the discrimination.

²Ibid., p. 9.

³Ibid., p. 88.

Test D, Generalization Test, is intended to measure the student's ability to generalize at an appropriate level of certainty. It contains a number of statements with the first word omitted (e.g., _____ people are in favor of socialized medicine), and the student is asked to complete the statement by selecting one of five words ranging from "all" to "no." Half of the items deal with non-controversial and half with controversial social issues. The 1952 version omits this test.

Test E, Discrimination of Arguments, presumably measures whether "the subject can weigh arguments fairly, regardless of whether they uphold or attack his own positions."⁴ The student is presented with a question, e.g., Should refugees from religious and political persecution in other countries be granted admission to the United States?; the questions are followed by four arguments. The student judges whether each argument is pro or con and whether it is weak or strong. By counting the number of times the student chooses weak or strong, the test presumably reveals a "tendency to be uncritical or hypercritical." The items are chosen from the area of socio-economic controversy. This test is similar to test 5, Evaluation of Arguments, in the 1952 edition except that the pro-con discrimination is not called for. The directions given to the student provide an example and explain in detail how the discrimination is made.

Test F, Evaluation of Arguments, is intended to measure

"appreciation of four principles: (1) If certain premises are granted then valid inferences drawn from these premises must follow. (2) Crucial words and phrases must be defined, and changed definitions will produce a changed conclusion. (3) The validity of an indirect argument depends upon whether the range of possibilities have been considered. (4) A logical argument cannot be disproved by ridiculing the arguer or his argument, or by attacking his motives, etc."⁵

The test consists of a short paragraph. For example:

In the course of a letter to a newspaper the president of a power and light company said: "The Federal government is building electric power lines which will compete with power utilities in the Tennessee Valley. It is unfair for the Federal government to compete with private power utilities."

The student is to assume the statements to be true and to choose one of three conclusions which are "consistent with them, and can logically be drawn."⁶

⁴Ibid., p. 91.

⁵Ibid., p. 92.

⁶Ibid., p. 200.

Six statements follow the conclusions and the student is to pick the one which explains why the conclusion is correct.

The parallel test in the 1952 edition of the Appraisal is called Interpretation. The student is given a paragraph and asked to judge whether each of a number of subsequent statements or conclusions "follows" or "does not follow." The items deal with equal numbers of controversial social problems and non-controversial problems.

Test 2, Recognition of Assumptions, which is included in the 1952 edition, but not in the original version, is intended to measure whether a student can recognize assumptions. The test consists of a number of statements each followed by three assumptions possibly underlying the statement. The student determines whether each proposed assumption is "made" or "not made." Half of the items deal with controversial social issues, and half do not. The directions for this section explain what an assumption is, and provide an example which explains how the authors think the student should arrive at a correct answer.

General comments on the Watson-Glaser tests. Glaser states that the goal of his study is to develop citizens who can think critically. He does not, however, indicate how the skills he has chosen were derived. There are any number of operations which would seem relevant but have been omitted, e.g., constructing hypotheses and evaluating authorities, and recognizing value dilemmas.

Perhaps Glaser considers these less important, but without reasons for the inclusion and exclusion of certain skills, it is difficult to evaluate Glaser's concept of critical thinking and his instrument. This leap from a general objective to the selection of a very limited number of analytic operations is particularly evident in Glaser's treatment of the problem of ethical conflict. Although he says, for example, that commitment to basic democratic tenets is extremely important, he has not chosen to include any discussion of the relationship of these values to the process of political analysis. Glaser's list of operations seems to constitute what may loosely be termed "scientific method," i.e., identifying problems, stating hypotheses, gathering evidence, and drawing conclusions. There are at least two difficulties here with the simple scientific model. First, the relationships within each category of operations and among the several categories are not shown. For instance, Test B is a measure of syllogistic reasoning. Test C is a measure of the ability to draw inferences. In what way these skills may be related to one another is not indicated. Second, even if the operations were presented in some systematic way, there is the question of whether the generalized "scientific model" provides the most appropriate rhetoric for dealing with the decision-making process in the arena of political conflict. From our point of view the natural science model or even the empirical social science model is a subsidiary component in the resolution or clarification of political controversy. The central problem is the definition and clarification of values, as well as mutual persuasion among antagonists that one or another priorities of goals should be embraced. In such an analysis the relationship

between facts, hypotheses, and generalizations and the general values of the culture is critical; and either taken in isolation is somewhat empty. Subtests A1 and A2 were an attempt to deal with values in isolation, and were finally abandoned. Subtests measuring valid inferences, logical reasoning, and ability to generalize at a correct level of inference handle empirical and logical elements in the argument without reference to values. The most critical element of all, however, the relationship between fact and value, is left out.

Eight Year Study of the Progressive Education Association⁷

The educational goals of this study were determined by asking thirty cooperating schools to list important educational objectives. These were collected and categorized by the evaluation staff into major types of objectives. In developing measures the usual procedure was to form a committee for each major category of objectives which then attempted to define objectives behaviorally and develop appropriate measures. The measures which fall within the scope of our work were developed by the Evaluation of Aspects of Thinking Committee and the Committee on the Evaluation of Social Sensitivity.

Measures developed by the Aspects of Thinking Committee. One of the justifications offered for tests developed by this committee was that they would serve to evaluate the attainment of a major educational goal in citizenship education.

The responsibility for training citizens who can think clearly has been so long and so frequently acknowledged that it is now almost taken for granted.⁸

"Clear thinking," the authors point out, is fundamental to all aspects of the educational program. Proceeding from this basis, the committee analyzed critical thinking into several components: (1) Ability to interpret data, (2) Ability to apply principles of science and social studies; (3) Ability to use logical reasoning in general; (4) Certain abilities associated with the nature of proof. The analysis of each of these abilities is discussed in some detail and a measure was constructed for each. (The ability to apply principles in the social studies was analyzed and evaluated by the Committee on Social Sensitivity.) Each of the tests constructed is approximately equal in length to all the sections of the Watson-Glaser appraisal combined.

The items for the Interpretation of Data Test were constructed on the basis of two considerations: The operations a student must perform in interpreting data, and the kinds of data the student should be able to interpret.

⁷Eugene R. Smith and Ralph W. Tyler, Appraising and Recording Student Progress, Vol. III of Adventure in American Education (New York: Harper and Brothers, 1942), p. 19.

⁸Ibid., p. 35.

Two operations were selected as paramount: perceiving relationships, and recognizing the limitations of data. Each in turn was defined more specifically. The ability to "perceive relationships" was described as the ability to make comparisons, to see common elements in data, and to recognize prevailing tendencies or trends in data. The ability to recognize the limitations of data involves recognizing what information in addition to that given is necessary to be reasonably certain that the interpretations and judgments about causes and effects are correct. The data used for the items were drawn from a variety of topics, e.g., data showing hourly wages in the automobile industry and the average cost of automobiles per pound; data showing the total number of union members in each of fifteen industries. In each item, following the presentation of the data in graphical or tabular form, fifteen interpretations of the data were given. The student is asked to indicate whether each statement is true, probably true, probably false, false, or whether there is insufficient data to make a judgment. Scores are derived for general accuracy (number of correct answers), excessive caution (number of "probably true" or "probably false" choices when true or false would be more appropriate), going beyond data (number of "true" or "false" choices when not justified).

The Logical Reasoning Test is an attempt to assess the student's ability to apply selected principles of logical reasoning to specific problems. Four principles of logical reasoning are emphasized in the test.

Definition: crucial words and phrases must be precisely defined, and a changed definition may produce a changed conclusion although the argument from each definition is logical.
Indirect argument: the validity of an indirect argument depends upon whether all the possibilities have been considered....

Argumentum ad Hominem: An attack upon an individual is not considered sufficient to prove that an individual or institution lacks merit. If, then: if one accepts the premises, then one must accept the conclusions which follow.⁹

The test presents problems. For example:

In 1940, W. Gibson Carey, Jr. then president of the Chamber of Commerce of the United States made the following statements in an address delivered at Waukegan, Illinois:

Our primary job is to kill excess government spending lest it kill us. Excess spending is inevitable when the central government takes over many of the duties which rightly belong to the states. The central government of the United States has taken over many of the duties which rightly belong to the states.¹⁰

⁹Educational Testing Service, Logical Reasoning Test Manual (Princeton: Cooperative Test Division, 1950), pp. 3-4.

¹⁰Educational Testing Service, Interpretation of Data Test, Upper Level (Princeton: Cooperative Test Division, 1950), p. 7.

The problem is followed by three possible conclusions from which the student chooses one. These conclusions are then followed by twelve proposed explanations of why the conclusion drawn is logical. The student chooses those statements which he thinks explain why his conclusion is logical; those which do not explain why his conclusion is logical; those about which he is unable to decide. The problems in the test deal with social security, highway building, labor, and government spending.

Interpretation of the test scores is extremely difficult since the test does not yield a single score which can be interpreted as an index of ability to think logically. The test provides a pattern of scores which the authors claim can be used for diagnosis. The limitations of the instrument for research are indicated in the test manual by this statement: "The number of patterns of behavior is almost as great as the number of students who take the test. Each pattern should be considered as a unique situation to be interpreted."¹¹

The Nature of Proof Test¹² attempts to measure a student's ability to analyze "proofs" of two kinds; those which proceed from stated convictions and those which are based in part upon data derived from experiments. The behavior which is involved is that evidencing "the ability to locate and appraise the basic assumptions upon which the soundness of the proof depends."¹³

The test consists of problems presented in paragraph form. One of the problems, for example, describes an experiment that deals with resistance of several brands of hand creams to water. The last line of the paragraph is an underlined conclusion. Following the paragraph are fourteen statements. "Some of these statements are relevant in the sense that they describe assumptions underlying the argument and some of them are irrelevant."¹⁴ Some of the relevant statements support the conclusion and the others contradict it. In the first part of the test the student is asked to distinguish the statements which are relevant and to classify them as supporting or contradicting. In the second part the student is asked to choose those supporting statements he would challenge. In the third part the student is asked to choose the one of three given conclusions which is most consistent with his own analysis. In the fourth part the student is asked to select from a list of proposals, those which he thinks would help him solve the problem. In the fifth part

¹¹Educational Testing Service, Interpretation of Data Test Manual, p. 12.

¹²Educational Testing Service, Logical Reasoning Test (Princeton: Cooperative Test Division, 1950).

¹³Eugene R. Smith and Ralph W. Tyler, Appraising and Recording Student Progress, pp. 127-128.

¹⁴Ibid., p. 139.

the student is asked to select those proposals from part four which he thinks he or his class could carry out.

The scores given to the student purportedly reflect "their success or failure in carrying out each of the procedures in the five parts of the test."¹⁵ This would seem to indicate that there are right and wrong answers, but then determination apparently cannot be objectively made.

The interpretation of the results depends upon the interpreter's understanding of the structure of the test problems. The usefulness of the test results is in direct proportion to the extent of the interpreter's concern with the objective and his confidence that significant behaviors involved in the objective are actually sampled in the different parts of the test. the number of patterns of behavior revealed by the test scores is almost as great as the number of students who take the test. Each pattern should be considered a unique situation to be interpreted.¹⁶

Measures developed by the Committee on the Evaluation of Social Sensitivity. Involved in the notion of social sensitivity, the authors say, are "many things."

Personal sympathies and aversions largely determine the pattern of initial awareness. The knowledge one possesses and the attitudes and viewpoints one has determine how one interprets the experience. One's ability and inclination to relate and re-organize ideas gained from previous experiences and to apply them to the new situation add insight. The inclination and ability to relate the feelings evoked and positions taken in specific situations to more general and abstract ideas add to both the coherence and depth of one's insights in a given case.¹⁷

The vagueness of this statement makes it difficult to understand whether the authors would include in the concept of social sensitivity the thought processes involved in the making of decisions about social issues. However, the committee in attempting to define operationally social sensitivity arrived at categories of behavior, one of which is:

Social thinking; e.g., the ability to (a) get significant meaning from social facts, (b) apply facts and generalizations

¹⁵Ibid., p. 141.

¹⁶Ibid., p. 43.

¹⁷Ibid., p. 158.

to new problems, (c) respond critically and discriminatingly to ideas and arguments.¹⁸

The Committee developed measures of "the ability to apply social generalizations and facts to social problems." In constructing such measures, the Committee assembled a list of fundamental generalizations and a list of behaviors which were necessary in applying these generalizations. It is interesting to note that in the process of trying to select generalizations, the Committee got embroiled in the question of the nature of the social sciences. There was considerable concern over the distinction between a social fact and generalization, and the truth or falsity of these generalizations.

Many of these generalizations seemed to apply only to a limited range of situations and lacked the universality commonly attributed to a "principle" Still other generalizations seemed to have little validity independent of a particular social philosophy or theory.... It seemed clear that the majority of useful and significant social science generalizations were not verifiable in the same sense as are the majority of scientific principles.¹⁹

The Committee classified the principles into types "each type perhaps implying a different technique for evaluation." The first type included descriptive generalizations serving merely to summarize a body of facts. The second type was explanatory generalizations and the third, value judgments, opinions and beliefs.

The behaviors involved in applying the facts and social generalizations were: (1) Relevance: Are the generalizations the student uses relevant to the situation? (2) Comprehensiveness: Does the student see the implications of generalizations and facts? Has he failed to consider the important generalizations? (3) Consistency: In using value or attitudinal principles is he consistent in the point of view he accepts? Does he use some principles which are conflicting either with each other or with the decision? (4) Objectivity and tenability: Does the student rely on generalizations which can be supported by fact or is he susceptible to slogans and loaded language?

Two instruments were developed; the first was called the Social Problems Test and presumably measured "the application of value principles or democratic tenets;" the second was called the Application of Social Facts and Generalizations and was designed "to appraise the application of facts and explanatory generalizations."

The "Social Problems Test" consists of exercises, one of which, for example, describes the sicknesses many workers develop as the result of poor

¹⁸Ibid., p. 161.

¹⁹Ibid., p. 170.

working conditions. Very few states, the paragraph points out, have adequate health regulations. The students are asked to choose one of three decisions regarding the problem of improved health regulations: Should they be left to the factory owners?; Should the government regulate?; Should a committee of workers and owners work out the problem? Following the decisions are twenty reasons that imply a value position. Each of the reasons is classified by the authors as democratic, undemocratic, or rationalizations. Democratic represents "defense of the interests of the general public or general welfare, of such democratic rights as freedom of speech, equality of opportunity, and a decent standard of living, of rights of minority groups"; while undemocratic represents "protection of special privilege, supremacy of efficiency and economic gain over human needs and values, undemocratic procedures or discrimination." Compromise represents "an effort to reconcile the two," and rationalizations are undemocratic values stated as democratic slogans.²⁰ An example of a rationalization is "Even though it is important to improve working conditions, it is undemocratic to accomplish this through dictation by the government."²¹ The student is asked to choose the "reasons" which are consistent with his decision. Each of the paragraphs in the test deals with larger social issues, e.g., conservation, free speech, socialized medicine.

The scoring and interpretation of this test is exceedingly complex. No single score may be interpreted as success or failure of the ability to analyze social problems or its components. "Each of the single scores describes only one part of a pattern; it is not justifiable to draw conclusions about an individual without considering the whole pattern of scores."²² There are two groups of patterns, one representing the logical aspects: comprehensiveness, consistency, tenability, relevance; the other representing a pattern of values: democratic, undemocratic, or rationalizations. The three scores used for reliability and validity are comprehensiveness, "the average number of reasons the student has marked for each course of action,"²³ consistency, "the number of reasons checked by a student which are inconsistent with the course of action chosen,"²⁴ and democratic values, "the number of times a student has used reasons expressing the values of general welfare and democratic rights."²⁵

The Applying Social Facts and Generalizations to Social Problems Test is intended to measure: (1) The ability to see logical relationships between

²⁰Ibid., p. 183.

²¹Ibid., p. 182.

²²Ibid., p. 188.

²³Ibid., p. 184.

²⁴Ibid., p. 186.

²⁵Ibid., p. 187.

principles and specific information and between the specific information and the social problem. (2) The ability to evaluate arguments in a specific social problem and discriminate between statements of verifiable facts, statements of opinion, and common misconceptions. (3) The ability to judge the consistency of policies with certain social values.

The test consists of nine problems in paragraph form. For example, one describes housing as a major problem and suggests two solutions: government regulation of building standards or government construction of housing. It goes on to state "Each method has certain advantages and disadvantages. Nevertheless, many people believe that the government should build houses at public expense to rent to those sections of the population with the lowest incomes." Twenty statements, either supporting, contradicting or irrelevant to the underlined solution are presented. The student is asked to decide which of the statements support, contradict or are irrelevant to the initial statement. In the second part of the test the student must indicate which of the statements can be proven true or false. There the student's responses are evaluated by the ability to discriminate facts from value assumptions. In the third part of the test the student is given a statement of a particular social objective (e.g., improvement of housing conditions), and several proposals for reaching this objective. His responses are evaluated by the number of times he chooses policies which are appropriate for achieving the objective.

There is no discussion of the details for interpreting the test results. The problems included in the test are consumer buying, health, unemployment, housing, soil conservation, civil liberties, international relations, taxation and civil services.

General comments on the Eight Year Study Tests. The Aspects of Thinking Committee, after justifying the measures in terms of citizenship education, provides an extensive catalogue of the specific abilities involved in "clear thinking." Clear thinking is apparently conceived of as a general ability which is of use in all areas, since no distinction is made, for example, between the abilities which are necessary for clear thinking about the physical sciences and social issues. As in the Watson-Glaser test, there is no discussion of how the set of thinking skills were derived, or of the relationship between these skills and the goal of good citizenship. There is no consideration of the central position of values in making political decisions or of the relationship between values and the thinking process.

The Social Sensitivity Committee, while it also fails to justify the selection of skills, does recognize the central position of values. They suggest that behaviors involved in evaluating "value judgments, opinions and beliefs" may imply a "different technique for evaluation." Yet the instrument for evaluating this technique, the Social Problems Test, offers no way of determining success or failure in dealing with value issues. The only score the test yields on values is a "consistency" score. The fact that consistency alone is considered a valuable trait, which is clearly implied, reveals, from our point of view, a basic misunderstanding of the nature and use of values in making social decisions. (See Parts One and Two.)

Both committees make some effort to define operationally "clear thinking" and "social thinking." As with Watson-Glaser, however, there is no explanation of the relationships among or within the categories of operations. The Aspects of Thinking Committee recognized this and offers this explanation.

In connection with any given problem, the process of reflective thinking, as defined by Dewey and others, is likely to call upon a number of the abilities to be described in connection with each major aspect of thinking mentioned above. It should also be noted that other important aspects of thinking -- for example, the ability to formulate hypotheses -- are only implicitly included in the above list and receive only cursory attention in the following discussions. The separation of clear thinking into these and other aspects is a product of the analysis and is not to be considered as inherent in the process of clear thinking. It was convenient because it facilitated the exploration of the larger objective and the development of a practicable means of evaluation. A satisfactory evaluation of the thinking abilities of the students involves a synthesis of the data obtained from various instruments.²⁶

Both committees' lists of operations constitute what may be called the "scientific model," although they appear in no explicit or systematic form. For instance, in the Interpretation of Data Test, the student's score is lowered if he goes "beyond the data;" in the Social Problems Test the student is called upon to be "objective." Even if this were presented systematically, we would again raise the question of the adequacy of this model. The Committee on Social Sensitivity recognized this problem when they reported that they became embroiled in the question of the nature of the social sciences in their attempt to collect a list of social science generalizations to be applied to social issues. However, they apparently made no progress on the question as far as we can determine from the report or the measures.

The Cooperative Study of Evaluation in General Education

The Cooperative Study was sponsored by the American Council on Education and was undertaken to evaluate the general education programs in twenty participating colleges and universities.²⁷ Of the several evaluation committees set up under this study, two are particularly relevant and will be discussed here: The Committee on Social Science Objectives, and The Committee on Critical Thinking. Both of these committees developed measures of critical thinking.

²⁶Ibid., p. 36.

²⁷L. Dressel and B. Mayhew, General Education: Explorations in Evaluation (Washington, D. C.: American Council on Education, 1954).

Measures developed by the Committee on Social Science Objectives. This committee selected the "development of traits incident to good citizenship" as one of the primary goals of education and undertook a study of the characteristics of good citizenship. The Committee's conception of "good citizenship" is described in the following statement.

A number of studies about the development of good citizenship characteristics [have been] noted. None of these has satisfactorily defined what good citizenship actually is. Defined one way, good citizenship consists merely of the practice of certain rudimentary mechanisms incident to democratic government. Such things as voting, knowing the provisions of the Constitution, and keeping abreast of current events are exemplary of this minimal definition. At the other pole lies the conception of citizenship as synonymous with the democratic way of life or the corpus of Christian ethics. Viewed in this latter light citizenship can no more legitimately be claimed as the exclusive province of the social sciences than of the natural sciences, the humanities, religion or any other aspect of college life.²⁸

The Committee goes on to say that if good citizenship were defined narrowly the only measure of achievement would be to count how many times graduates voted, attended town meetings and so on. Rejecting this, they say that conceiving citizenship in the broader sense would involve such a wide variety of tests of knowledge, attitudes, values, and beliefs that their development would be outside the scope of the Committee. But, the authors say, although these matters were beyond their scope, critical thinking, one of the essentials of good citizenship can be evaluated. "The essence of good citizenship is that each person possess the potentialities for discovering his own problems and for developing socially acceptable solutions to them."²⁹ The Committee then seeks to define critical thinking. It compiled what it describes as an "a priori" list. Critical thinking consists of the ability --

1. To identify central issues
2. To recognize underlying assumptions
3. To evaluate evidence or authority
 - (a) to recognize stereotypes and cliches
 - (b) to recognize bias and emotional factors in a presentation
 - (c) to distinguish between verifiable and unverifiable data
 - (d) to distinguish between relevant and non-relevant
 - (e) to distinguish between essential and incidental
 - (f) to recognize adequacy of data
 - (g) to determine whether facts support a generalization
 - (h) to check consistency
4. To draw warranted conclusions

²⁸Ibid., p. 35.

²⁹Ibid., pp. 35-36.

The measure. The Committee then proceeded to construct a single instrument to sample all these abilities in "The Test of Critical Thinking in Social Science."³⁰ The test consists of fifty short items which vary in format and the kind of operation called for. It is not clear from the report or the test, which items are intended to get at each of the abilities in the list. Some items are common reading comprehension items; some present the student with a paragraph in which he is asked to judge the relevance of several statements which follow; some items present the student with a series of statements which are a mixture of value judgments and empirical claims and the student is asked to distinguish between them; some items present statistics in tabular or graphical form and the student is asked to choose a generalization which follows. The test yields a single score; there are no sub-test scores.

Tests developed by the Committee on Critical Thinking. This Committee justified its concern for measuring critical thinking with the following statement.

The very existence of the Inter-college Committee on Critical Thinking at once bears witness to the generally accepted importance of the objective ³¹

Although it is not made explicit, there is a clear implication that the major goal of all general education is the teaching of critical thinking.

The Committee collected a list of critical thinking behaviors which they categorized under five major headings. (1) The ability to define a problem. (2) The ability to select pertinent information for the solution of a problem. (3) The ability to recognize stated and unstated assumptions. (4) The ability to formulate and select relevant and promising hypotheses. (5) The ability to draw conclusions.³²

There is a brief "rationale" for each of these behaviors which attempts to explain the importance of each operation. The authors say that these operations "interact." However, the "rationale" does not describe the nature of this interaction, but merely reasserts the importance of the operations.³³

³⁰American Council on Education, Cooperative Study of Evaluation in General Education, A Test of Critical Thinking in Social Science (Washington, D. C.: American Council on Education, 1951).

³¹L. Dressel and B. Mayhew, General Education: Explorations in Evaluation, p. 174.

³²Ibid., p. 179.

³³Ibid., pp. 179-181.

The actual instrument developed by the Committee, A Test of Critical Thinking, included items which were intended to sample the abilities listed above. The five types of items corresponding to these abilities are described below.

Ability to define a problem: A one or two line statement is followed by five statements of what the problem may be. For example:

The Kemp family wishes to repaint its livingroom walls. Their problem is

1. What color and kind of paint will best fit the family's use of the room and budget of time and money.
2. What color goes best with the rugs and curtains.
3. How best to time the painting in relation to the baby's sleep, Jane's birthday party, and other events scheduled for the house.
4. What kind of paint -- water or oil base, etc. -- is cheaper in the long run.
5. Whether they should use wallpaper since it will be cheaper and more colorful.³⁴

Ability to select pertinent information: A short paragraph describing a situation and including a conclusion is presented. The paragraph, for example, describes a town in which there was an epidemic of typhoid after some typhoid contaminated wastes had been thrown into a nearby river. The conclusion is given that the wastes caused the epidemic. The conclusion is followed by statements and the student decides whether each statement supports, infirms, or is irrelevant to the conclusion.

Ability to recognize assumptions: The student is presented with a paragraph which contains a number of claims; e.g., one paragraph is a newspaper advertisement for an aluminum duct describing its utility, economy and durability. Following the paragraph are ten statements and the student decides whether each (1) is stated in the paragraph, (2) is an unstated assumption, (3) is not stated and is irrelevant, or (4) is not stated, but would weaken the argument if included.

Ability to formulate and select relevant hypotheses: The student is given a "riddle type" item, e.g., "You are stranded in a waiting room. You are hungry. You find a vending machine about which you know nothing"

³⁴American Council on Education, Cooperative Study of Evaluation in General Education, A Test of Critical Thinking, Form G (Washington, D. C.: American Council on Education, 1952). The keyed answer is "1." Ennis criticizes the items in this part of the test, with specific reference to this item.

The student must select hypotheses to fit the riddle and, as more information is provided, test it out and select another if it does not hold.

Ability to draw valid conclusions and judge the validity of inferences: These items take two forms. The first presents a statement followed by several conclusions and the student chooses the most appropriate; the second presents a short problem which includes a conclusion and the student is asked to choose one among several explanations of why the conclusion is sound.

The five sections of the test yield one score. The content covered by the test is broad, ranging from vending machines and repairing automobiles to several which deal with controversial social issues.

General comments on the tests developed by the Cooperative Study in General Education. The Committee on Critical Thinking comes closest to suggesting a model of critical thinking, implying that the five skills they list are a related sequence. However, the relationships are not indicated, and the items in the test deal only with fragmented elements of these skills and not with the relationships among them. None of the items, for example, follows through a sequence from the identification of a problem to the formulation of a hypothesis, to the selection of evidence and the drawing of a conclusion. Like the Watson-Glaser and the Eight Year Study Test, A Test of Critical Thinking in the Social Sciences and A Test of Critical Thinking, Form G, are committed to what may loosely be termed scientific method. Again we would question the adequacy of this as an exclusive model for dealing with social and political issues.

The Illinois Critical Thinking Project. The Illinois Project, like the Glaser study, focused more on the problem of teaching critical thinking than on the problem of measurement. It was assumed that after a year of more or less systematic treatment in critical thinking students would score better on the Watson-Glaser Appraisal and the ACE Test of Critical Thinking, Form G. A more specialized test was drawn up by the Project staff, however, called A Test of Critical Thinking.³⁵ This test presumably measured the student's knowledge of principles of critical thinking.

While some preliminary findings of this study are reviewed in Chapters Eleven and Twelve, we should mention here the Illinois test along with a factor analytic study comparing it with Watson-Glaser and ACE Form G carried out by Rust, Jones, and Kaiser.

Rust, et. al. found that

Two of the three factors obtained are fairly strong; one seems quite appropriately to bear the label General Reasoning, and the other, Logical Discrimination or Application of Logical

³⁵B. O. Smith, and others, Teaching Critical Thinking (Urbana: University of Illinois Press, in press).

Principles. The third factor, however, is less clear. It may represent sophistication in the use of language and for this reason has been tentatively called Semantics or Verbal Understanding. On the other hand, it may only be measuring the relative difficulty of the items.³⁶

This study confirms our own suspicion that Watson-Glaser and the ACE Test get more at some general reasoning and/or intellectual factor, rather than specific critical thinking abilities. The most interesting finding, however, is that Smith's Principles test apparently does measure something else, especially the subtests labeled "Semantics I" and "Material Fallacies." An illustrative item from each of these subtests is reported in the Rust study.

Item from Subtest Semantics I:

"The apple is red, large, and juicy. Therefore it is a good apple."

This statement is to be marked as:

- "A -- the truth of the assertion can be decided with facts, either those given or additional facts.
- B -- the truth of the assertion is difficult to decide because criteria (standards) for deciding its truth are not generally agreed upon.
- C -- the truth of the assertion can be decided without facts."

Item from Subtest "Material Fallacies":

"Acids turn blue litmus paper red. All three of these liquids turned the litmus paper red, and hence they are acids."

This statement is to be marked as:

- "A -- it is a hasty generalization.
- B -- it begs the question.
- C -- it commits the fallacy of composition.
- D -- it commits the fallacy of argumentum ad hominem.
- E -- none of the above."

The obvious difference between the Illinois test and the other two tests, it would seem to us, is the amount of technical information required to do well on the test. As we have already noted, both Watson-Glaser and the ACE test are application tests in that they contain most of the information required to do the items within the tests themselves. The directions, in fact, are brief training procedures describing how to do the items. The Illinois test,

³⁶A. Rust, "A Factor Analytic Study of Critical Thinking," The Journal of Educational Research LV (March, 1962), p. 258.

on the other hand, assumes specialized training and therefore takes the liberty of using technical language and concepts without informing the person taking the test what the language or concepts mean. It is an application test only if one assumed that the person taking the test has knowledge of the abstract concepts which he is to apply. This assumption probably does not hold for many high school youngsters.

While Smith's strategy for test construction is certainly defensible in the light of his particular research problem the Illinois test has only limited value as an intermediate measure of knowledge of and simple application of critical thinking principles. We might report here, however, that one of the members of the Illinois Project, Robert H. Ennis, has continued work on the systematic definition of critical thinking³⁷ and the use of this comprehensive definition for the development of more sophisticated measures.³⁸ Unfortunately, Ennis has explicitly chosen not to deal with two of the most fundamental problems of measurement in this area: (a) handling the problem of clarifying values, and (b) dealing with the interrelationships among problems as they bear on a total argument. On this latter point he is concerned with assessing the truth value of specific statements rather than the power of a total argument.

The Iowa Test of Educational Development, No. 5, Interpretation -- Social Studies

In this general review of measurement efforts in the area of critical analysis of social issues, we should mention the Iowa Test, No. 5. Although it presents no striking departure from other efforts, it is in print and commercially available. It consists of a series of passages to be read, and a series of questions to be answered based on each passage. Probably the greatest weakness of the test is the uneven levels of "interpretation" required, which reflect even greater lack of clarity in its rationale than is evidenced by the other efforts mentioned. Items range across a variety of types of cognitive operations. These include requests for simple reading comprehension:

In which part of the United States were the first steps taken toward improving the care of the poor and the insane? (The answer is contained in the passage read by the student.)

Requests for simple recall of memorized information:

Which of the following was a worker of the same period whose

³⁷Robert H. Ennis, "A Concept of Critical Thinking," Harvard Educational Review, XXXII (Winter, 1962), pp. 81-111.

³⁸R. H. Ennis and J. Millman, Cornell Critical Thinking Test (Mimeographed).

achievements in another field resemble those of Dorothea Dix?
(The answer is not contained in the passage read by the students.)

Requests that the student evaluate suggested implications from available data:

Writer V implies that some Congressmen (1) like to fool the public, (2) work too hard, etc.

Requests that the student analyze the proof process:

Which, if any, of the writers supports his statements by specific facts and examples?

Request for inferences:

Adams' policy in this affair is best described as (1) traditional (2) aggressive, etc.

Like the Test in Critical Thinking in Social Science, the items are, in fact, so heterogeneous, that it is difficult to know precisely what the authors who constructed the test had in mind, except perhaps some general notion of an ability to read intelligently and critically. One strength of the test, however, is that it is apparently not tied to a very limited number of explicit cognitive operations, or to a simple model of "problem solving." Since the authors are not wedded to these simpler notions, however, they attacked the problem intuitively rather than systematically, which makes the test extremely difficult to interpret.

Problems with Available Measures

1. The problem of validation. None of the measures discussed above have been validated in the sense that it has been determined whether or not those who score high in these instruments also can think systematically and rigorously about social issues in situations other than those provided by the test. (Smith, in fact, reports that he was surprised at the low correlation among a wide variety of critical thinking subtests.) Can we predict, for example, that a person will handle a controversial discussion situation or an editorial more competently because he has done well on one of these tests? There is presently little evidence on this point.

2. The leap from citizenship to critical thinking. The authors of the Watson-Glaser appraisal, the Eight Year Study Evaluation Procedure, and the General Education Tests all imply a relationship between "good citizenship" and critical thinking. If we assume that citizens should think critically in making political decisions, then it seems to follow that some basic questions about the components of the decision-making process should be raised. In other words, what things must an individual consider in arriving at an intelligent decision? A major omission in the tests reviewed is their failure to deal with ethical issues. Political decisions in the western democracies are

made within the context of western political values.³⁹ These values are not only more or less common to all Americans but they often conflict. The decision-making process, then, must involve some strategy for dealing with these conflicting values. The implications of this point of view for curriculum are developed in the earlier chapters. The point here is that the development of measures of critical thought about social issues which are by their nature controversial must also consider the political tradition within which the society operates. The mere listing of analytic operations without showing how they are to be used in a controversial situation is a fundamental difficulty of all the measures thus far reviewed. This, of course, raises the question of the adequacy of the "scientific method" itself for dealing with social issues.

It is not our intention here to raise the question of what scientific method is and to what extent the tests measure its essential characteristics. However, we do wish to ask whether or not the scientific model adequately deals with questions of social policy. Beard points out characteristics of the social sciences that are particularly relevant to this point.

Nevertheless, the social sciences can furnish knowledge, an insight into the processes of social thought, and a method indispensable to the operation of formulating workable objectives. They disclose, by examination and tabulation, the objectives which American society, acting through the public agencies of education and private associations, has already proclaimed. They present facts and organizations of knowledge and thought corresponding, with more or less accuracy, to the realities of American society and world relations which must enter into the formulation of any workable scheme of objectives. They describe the general trends or tendencies of society which provide the only available indications for forecasting the necessities and probabilities of the future social situations in which children now in the schools will have to live their lives and discharge their obligations. They describe the ethical and esthetic ideals which furnish guidance in determining what is desirable within the limits of necessity, which may be potentially realized, in whole or in part, and which are bidding for the loyalty of coming generations. They describe qualities of human character that represent the ideal and are necessary to the maintenance and development of a society compatible with the realization of the ideal.⁴⁰

³⁹Carl J. Friedrich, *Constitutional Government and Democracy* (Boston: Ginn and Co., 1950), p. 160; Gunnar Myrdal, *An American Dilemma* (New York: Harper and Brothers, 1944), chapter 1.

⁴⁰Charles A. Beard, *The Nature of the Social Sciences* (New York: Charles Scribner's Sons, 1932), pp. 180-181.

It is Beard's conclusion, however, that while the social sciences can be of assistance in making ethical choices, they cannot tell us what is right or wrong or what should be done to make an ethical decision.

Cohen and Nagel discuss the difference between what is and what ought to be as follows:

If we ask whether all men do respect human life, it is relevant to point out that some do not. But if we raise the issue as to whether we ought to or should refrain from murder, the mere fact that some men do not is not directly relevant. The validity of a moral command is not denied by the fact that some do not follow it. Moral maxims, as imperatives concerning what ought to be, differ from natural laws as uniformities of existence. The ordinary man attending to his daily affairs and a good many philosophers who call themselves positivists are agreed not only that there is this difference, but that there is no science of moral imperatives.⁴¹

C. I. Stevenson, John Hospers, E. C. Ewing, and Bertrand Russell,⁴² though they disagree over the process of arriving at ethical decisions, do draw a distinction between ethical and non-ethical statements. These writers generally maintain that the "scientific method" is in some measure useful in arriving at ethical decisions, but they hold that this method alone is not sufficient. We are not implying that there is no disagreement on this point in ethical theory, but at the very least it is a controversial point which would require that any position an author of a critical thinking test adopts should be explicitly stated. With the exception of Ennis this was generally not done by the authors of the tests that have been reviewed.

3. The problem of the fragmented nature of the test format. The general issue we are raising here is an ambiguity in the conception of the analytic operations one wishes to assess in a particular test. One can conceive of these operations as elements in a broader context, as those involved in "rational problem solving." Or one can conceive of these operations as embedded in general strategies for approaching political controversy. Or one can conceive of these operations as specific, but important, fragments of "critical thinking." This ambiguity can, perhaps, best be illustrated by two

⁴¹Morris R. Cohen and Ernest Nagel, An Introduction to Logic and Scientific Method (New York: Harcourt Brace and Co., 1934), p. 363.

⁴²E. C. Ewing, "Subjectivism and Naturalism in Ethics," Readings in Ethical Theory, ed. by W. E. Sellers and John Hospers (New York: Appleton-Century-Crofts, 1952), p. 120; John Hospers, An Introduction to Philosophical Analysis (New York: Prentice-Hall, 1953), p. 494; Bertrand Russell, "The Elements of Ethics," Readings in Ethical Theory ed. by W. E. Sellers and John Hospers (New York: Appleton-Century-Crofts, 1952), p. 8; Charles L. Stevenson, Ethics and Language (New Haven: Yale University Press, 1944), pp. 113-114.

definitions given in the literature. Ennis defines critical thinking as "the correct assessing of statements."⁴³ Bloom, however, uses the term more broadly and focuses on the analysis of a problem situation, rather than on the criteria for correctly assessing statements.

....It is also expected that students will acquire generalized techniques for dealing with new problems and new materials. Thus, it is expected that where the student encounters a new problem or situation, he will select an appropriate technique for attacking it and will bring to bear the necessary information, both facts and principles. This has been labeled "critical thinking" by some, "reflective thinking" by Dewey and others.... The most general operational definition of these abilities and skills is that the individual can find appropriate information and techniques in his previous experience to bring to bear on new problems and situations....⁴⁴

It is probably safe to assume that the efforts to evaluate competence in critical thinking reviewed above have concentrated on wholistic problem analysis. Because of this fact, we might ask whether the short item multiple choice type instrument, organized into subtests, each of which measures a separate analytic function, is the most appropriate way to assess what are essentially the broader problem analysis skills.

For example, the Watson-Glaser Appraisal includes an inference test in which each item in the test is intended to measure a student's ability to evaluate inferences. In the General Education Test of Critical Thinking there are some fifteen items designed to evaluate the student's ability to recognize unstated assumptions. The objection to this procedure may be seen more clearly if we look at a transcript of an actual discussion. This discussion was held by a group of eighth grade students on the problem of public school desegregation.

Doris: I'd like to tell the people who say the Supreme Court does not have a right to tell us to integrate the schools: it is a fact that you're not giving the Negroes their constitutional rights. That's why they're telling you to integrate. It's not that they're trying to interfere with your school system, but the Negroes don't get their constitutional rights, their natural rights. The Supreme Court can rule on this.

Jean: The Supreme Court has no right to tell us what to do; it's a state right; and I believe the schools are run right.

⁴³Robert H. Ennis, "A Concept of Critical Thinking," p. 83.

⁴⁴Benjamin S. Bloom, Taxonomy of Educational Objectives (New York: Longmans Green, 1956), p. 38.

Doris: But what about the constitutional rights, their natural rights? They're not getting an equal schooling.

Jean: They have the same schools; they have good schools.

Doris: They do not have good schools.

Bill: Right here it says segregation harms the Negroes by making them feel inferior. It doesn't matter whether the schools are equal or not. Segregation always makes the children feel inferior --

Jean: I don't think the Supreme Court has any right to make any such decision or any such generalization. Our schools for the Negroes are better than the schools for the whites. You could reverse that and say that Negroes feel superior to the whites because they have better schools. ⁴⁵

Note that the third time Doris speaks she has an opportunity to question the evidence for the prior assertion, yet she does not. Now, it may be that Doris knows how to question evidence but does not know when. Note also how Bill, in effect, claims that the issue of how good the Negro schools are is irrelevant, but Jane misses Bill's meaning. It may be that Jane has the ability to "identify the problem" but does not recognize that it could be of some assistance here.

Doris and Bill could perhaps score very high on "questioning evidence," and "identifying problems" respectively. The tests, however, may not evaluate the ability of the student to use the operations outside of the highly structured test situation. Competence in identifying a problem, questioning assumptions, and so forth, would seem to be academic if the student cannot use them in the course of dealing with a social issue in a more realistic context. It may be possible to construct a test which would measure this "when" component but it would probably require longer items which give the student the opportunity to select among several operations. Moreover, some tests not only structure the item so that the choice of operations is restricted but go on to teach the student how to perform the operations before he takes the test.⁴⁶

Our point is simply that while some of the instruments attempt to measure various fragments of a critical thinking process, none of them assess the ability to follow a sequence of operations within a dialectical framework, e.g., the sequence of stating a generalization, questioning a generalization

⁴⁵Donald W. Oliver, "Educating for Responsible Individualism," 30th Year-book of the National Council for the Social Studies ed. by Franklin Patterson (Washington, D. C.: National Council for the Social Studies, 1960), p. 224.

⁴⁶The Watson-Glaser is particularly guilty in this regard.

by providing a contradictory example, qualifying a generalization to take into account the contradictory example, and supporting it with authoritative evidence, including the credentials of the authority. One can see in the efforts of the PEA group attempts to get at this. The more complex the test becomes, however, especially within the multiple choice framework, the more difficult it is to develop an instrument in which the complexity of the directions or the ambiguity of interpretation does not substantially reduce its usefulness.

AN APPROACH TO THE EVALUATION OF COMPETENCE IN THE ANALYSIS OF POLITICAL CONTROVERSY

Suggested Criteria for Test Construction

In our own efforts to evaluate competence in analyzing political controversy, we have tried to avoid some of the pitfalls and difficulties described above. Although we have not always succeeded, we do feel that progress has been made. We suggest below some of the necessary steps for such progress.

1. The nature of the problem to be analyzed by the student has to be more narrowly construed than has thus far ordinarily been the case. To develop a general model as an analytical basis for a wide diversity of problems is too ambitious. Surely the important distinctions between problems which are essentially ethical, scientific, or aesthetic, for example, should be taken into account. Our work has focused on the analysis of the ethical component (as opposed to the political process component) in political controversy.

2. The situational context for measuring analytic skills should be described more carefully and the testing situation broadened beyond the simple multiple choice pencil and paper test. A number of elements should be considered in describing the test situation: (1) the form of the message with which the student is required to deal; (2) the content of the message which is the subject of analysis; (3) the extent to which the student is required only to analyze an existing message in terms of analytic concepts taught as against using these analytic concepts to evaluate or construct counterarguments; (4) the extent to which choices are prestructured for the student as against requiring the student to structure his own answer; and (5) the extent to which oral interaction is an element in the evaluative situation. These five considerations suggest the following kinds of choices. (These choices are obviously not exhaustive.)

a. form of message:

one-sided persuasive communication v. two-sided dialogue
(e.g., newspaper editorial or
advertisement.)

b. content of message:

scientific decisions v. public policy decisions v. aesthetic decisions

c. level of analytic abstraction

responses required in test might consist of abstract analysis of controversial dialogue v. responses might consist of appropriate rebuttal statements

d. level of structure

responses required in test situation might be prestructured and simply chosen by student v. test might require that student create the response

e. interpersonal context

student might be operating within a live interaction situation (e.g., an oral examination) v. student might be dealing with a controversy only as a non-participant or observer

In our work in test development we made the following decisions:

- (1) to use the dialogue as the basic form of message;
- (2) to deal with public policy decisions as the content of the message;
- (3) to develop measures concerned both with abstract analyses of the controversial dialogue as well as measures requiring the student to develop appropriate rebuttal statements constructed on his own initiative;
- (4) to develop both prestructured tests and open-ended unstructured tests;
- (5) to develop measures in which the student both analyzes the arguments of others and participates in the give-and-take of an argument himself.

These decisions resulted in four measures which are discussed in some detail below.

3. The analytic operations one wishes to evaluate must be described systematically and the student must be required to identify not only when the operation has been carried out successfully, but also when the operation is appropriate or relevant in a problem context. Systematic efforts have been made along these lines by several investigators including Ennis, Dressel

and Mayhew, and in the PEA studies⁴⁷ discussed earlier in this chapter. The major problem with these efforts, we feel, is the isolation of the analytic operation from the total problem context. Although there is a general concern with problem analysis, all choose to measure fragments of the analytic process isolated from other components of the process. This often results in building systems of thought to think about thinking which are so unwieldy and complex, or, at times, useless, that one scarcely has time to think about the problems they are meant to clarify.

Major intellectual operations assessed in Project measures. Our own approach has been much simpler than those used or suggested by Ennis or Smith and more "wholistic" and interrelated than the work of Dressel and Mayhew. We concentrated on five areas of analysis:

- a. problem identification and differentiation;
- b. making explicit important cross problem assumptions;
- c. identifying and using appropriate strategies for dealing with different types of problems;
- d. identifying common dialectical operations; and,
- e. identifying relevance problems.

Each of these areas of analysis is related to the others, as should be clear from the subsequent discussion.

a. Problem identification and differentiation. We were interested in the extent to which the student could differentiate among various classes of problems within a controversial setting. For our purposes we differentiated between empirical controversy, value controversy, and definitional or analytic controversy. Evidence of the student's ability to make these distinctions can be obtained in at least two ways: (1) he correctly labels or describes the nature of the problem, or (2) he uses a correct strategy for dealing with the controversy. For example, the student may say that a particular statement presents a problem over the meaning of a word, or it presents a problem concerning which of two conflicting factual claims is true. Or the student may say that we deal with this kind of problem by finding out how the word is commonly used in this context, while we deal with that kind of problem by looking for consistencies and inconsistencies in the reports of the event, investigating the credentials of the observers, etc. The use of an incorrect strategy can be illustrated by the following example:

⁴⁷Robert H. Ennis, "A Concept of Critical Thinking," pp. 81-111; L. Dressel and B. Mayhew, General Education: Explorations in Evaluation; Eugene R. Smith and Ralph W. Tyler, Appraising and Recording Student Progress.

Two men are arguing over whether the absence of fire escapes from two-story buildings constitutes a menace to public health. One says it is a menace; the other says it is not. The statistics regarding deaths caused by the absence of such fire escapes are available, and are not brought into the debate.

In this kind of situation students commonly state that the way to settle the argument is to go out and observe buildings without fire escapes or observe the statistics and see whether or not such buildings constitute a menace. From our point of view, one must observe not only the statistics and the buildings, but how the word "menace" is commonly used.

b. Making explicit important cross-problem assumptions. The categorization of a particular problem as definitional, factual, or ethical is only the initial step in a complex process of justification. For every problem that is identified, one must make countless assumptions about agreement (or question whether or not there is agreement) regarding related problems inherent in the same issue. Suppose one says, for example, "The Federal Trade Commission is needlessly censoring television advertising." The antagonist says "no!" If an argument ensues, it can take the form of a definitional disagreement (perhaps over the word "censorship"), a factual disagreement (over whether or not such censorship actually takes place), or a value disagreement (over whether or not the government should interfere with the free flow of information over the airways.) Each of these disagreements tends, however, to make assumptions about agreement in other areas, which may, in fact, be questioned. The value disagreement, in this case, for example, may assume that censorship is a real possibility or that it has actually taken place -- it assumes a fact. The factual disagreement generally assumes that there are common values both with respect to the free flow of information and to government regulation or control in the interests of the welfare of the community; otherwise the fact would not be worth arguing about. The definitional argument may also assume that censorship is imminent or is taking place, and that censorship can be bad. Analyzing important empirical and definitional cross-problem assumptions is especially important in the process of argumentation, since one of the subtlest ways to press one's value commitments with respect to a given controversial situation is to distort or exaggerate the information describing the situation, or to choose vague but loaded words to describe it.

c. The process of identifying and using appropriate strategies for dealing with different types of problems. Both analytic processes discussed above involve essentially problem identification and differentiation. From observing a great many discussions, we noted that discussants tend to avoid dealing with the various aspects of an argument systematically and prefer rather to "go around in circles." The following example illustrates a common pattern of reasoning:

Southern schools shouldn't be segregated because
segregation of the races is bad; moreover,

the Negroes in the South are given poor treatment under segregated conditions;

therefore Southern schools shouldn't be segregated.

Analytically, the line of reasoning can be described as follows:

A policy decision is supported by a specific value judgment (which really adds little new information).

Both of these statements are followed by an unsupported and controversial generalization.

The case is then terminated with a restatement of the initial policy decision.

This pattern of thinking is usually an inefficient way to approach an issue. Disagreements over decisions, for example, may be clarified by construing the decision in terms of the general values of the culture, discovering analogous situations in which the same values are in conflict so that the individual can see the problem in a broader context, identifying differences among the analogous situations which cause one to change one's decision from one situation to another, and testing whether or not these presumed differences do, in fact, hold up under careful scrutiny. If the differences do not hold up, the individual can change his decision and simply accept his own inconsistency. Disagreements over fact can be approached by constructing testable hypotheses and looking for evidence. Definitional disagreements can be approached by a process of categorical reasoning (Do the essential criteria that define a category apply to an object, event, or person we wish to include within the category), by a process of empirically testing how a word or label is commonly used, by pragmatically testing whether the particular use of a word will avoid ambiguity and confusion, or simply by seeking stipulative agreement on the use of the word in this particular situation. The strategies suggested here are obviously not exhaustive. The point is that some strategies are appropriate for some kinds of problems and inappropriate for others. And more important, we assume that making strategies of problem resolution explicit tends to lead the discussant to greater focus and more systematic analysis than simply making casual and perhaps unrelated statements in an attempt to justify a decision.

d. Identifying and using common dialectical operations. In the process of analyzing political controversy we have noticed a pattern of thought which seems common to the three major types of disagreement. The essential operations in the pattern may be described as follows:

generalizing: assuming that what is good, true, or useful in specific instances is good, true, or useful in general.

specifying: supporting, contradicting, or simply elaborating a general statement by pointing to specific instances in

which the general statement holds or does not hold. The concept "specifying" includes, for us, then, the operation pointing out the consistency or inconsistency of a specific statement with a more general statement.

qualifying: generalizing in a qualified way so as to take into account exceptional instances in which the general statement appears to be inconsistent with related facts, values, or definitions.

The following example will illustrate:

Statement (A) "Desegregation will improve education for the Negroes in the South."

(The individual making the statement is assumed to know of specific instances which make him think this generalization will hold, or to have heard the general statement from some authoritative source, and to believe, therefore, that it is generally true. This is an example of a general statement, although only implicitly does it illustrate the process of generalizing.)

Statement (B) "Sure, just like education for the Negro improved the first year when Little Rock's Central High School was integrated."

(This statement provides a specific example of a situation in which the initial generalization is presumed not to hold. It illustrates the process: specification-inconsistency.)

Statement (C) "When a careful plan of community education has been carried out, and the community is ready for it, integrated schools provide a better education for the Negro and white alike."

(This is a reworking of statement (A), with a qualification added which presumably takes into account the inconsistent example specified in statement (B).) The process of going from specific to general (generalizing), then returning again to specific test examples of the general statement (specifying), and then restating the general statement to take into account inconsistencies and exceptional cases applies to all three types of problems -- value problems, factual problems, and definitional problems. These concepts give us, then, an abstract conceptualization of dialectical strategy which cuts across problem types. The process of arriving at a qualified generalization, we think, is a very important aspect of the definition of reflective thinking.

e. The process of identifying relevance problems. Relevance can be viewed from two points of view: What statements are relevant to the argument as a whole; and what statements are relevant at a particular point in the discussion. Each discussion has a context set mainly by the topic. Some statements are clearly not relevant because, while they may be within the political-ethical frame, they are on a different topic, e.g., a statement about labor unions in a discussion of desegregation. Within the discussion itself some statements may be relevant to the discussion in general, but inappropriate to the immediate context. For example, the central problem of a discussion, at some point, may turn on what are in fact behavioral differences between whites and Negroes. Before this issue is in any sense settled, someone may move to the question of whether or not differences are culturally conditioned or genetically conditioned. While this issue may be relevant to the total discussion, it may be inappropriate at this particular point in the discussion, since it shortcircuits the immediate issue under analysis.

The Development of Four Experimental Measures

We shall now describe the results of our own efforts to evaluate competence in the analysis of political controversy. We developed four instruments, which actually constitute four strategies for getting at the same type of competence, but emphasizing different elements in the testing situation. All four tests are labeled the Social Issues Analysis Test (SIAT), with numbers designating different measurement procedures.

SIAT No. 1: Argument Analysis Test

SIAT No. 1 is a group administered pencil-and-paper test with precoded multiple choice responses. The purpose of the measure is to assess how well the student can identify the intellectual operations occurring in an argumentative dialogue. The items on the next two pages illustrate the form and nature of the test.

Both SIAT No. 1 and SIAT No. 2, which we shall describe in the next section, were originally composed and pretested on groups of Master of Arts in Teaching candidates in social science education at Harvard. We chose this sample in order to get some idea of whether or not the analytic framework of the Project had become so narrow or was expressed in such technical language that intelligent, well-educated lay people might not understand it. Since Harvard social science AMT candidates are generally students from the top quarter of graduating classes of well known liberal arts colleges, this seemed an appropriate group on which to run trial versions of our pencil-and-paper tests. Table 10.1 summarizes a typical first trial run. It seemed fairly clear to us that our language was not overtechnical nor was our conceptual framework idiosyncratic to the Project. The source of ambiguity in poor items was easily discernible and correctable. It was possible with one or two revisions of specific items to get almost complete reliability among adult judges. The reliability estimate for the SIAT No. 1 with a sample of

A Discussion about Labor Unions

Joe: Labor unions were O. K. when they first began. Now they are corrupt and dishonest. I say unions should be outlawed.

Mike: What makes you think they are corrupt and dishonest?

(l) Joe: Look what has happened to the Teamsters Union. A couple of years ago two of their biggest leaders were convicted of stealing union money.

NOW ANSWER QUESTION (9) AT THE RIGHT.

(j) Mike: Every now and then bankers are caught embezzling money from the bank. But no one suggests that banks be outlawed to get rid of dishonest bankers.

NOW ANSWER QUESTION (10) AT THE RIGHT.

Joe: That's different. Banks perform a necessary service. We couldn't get along without banks.

(k) Mike: Many workers feel that labor unions are necessary to protect their welfare, to see that they are not overworked and underpaid by irresponsible business leaders.

NOW ANSWER QUESTION (11) AT THE RIGHT.

(l) Joe: But once you let the labor union organize a factory, it tries to dictate how to run the whole factory: how much to pay the workers, when to give them holidays, how many hours they can work, who is first to be laid off when work is slack. Labor unions go against the whole idea of free enterprise and the right to make any kind of contract you please with your workers.

NOW ANSWER QUESTION (12) AT THE RIGHT.

Answer the following questions on the basis of the discussion at the left.

9. What does Joe's statement (i) do in the discussion?

- ☐ a. Gives examples to support a fact or claim.
- ☐ b. Gives examples to suggest that a definition is not adequate.
- ☐ c. Suggests that Mike's position in the argument needs qualifying.
- ☐ d. Gives examples to support or illustrate a definition.
- ☐ e. Clarifies a word or phrase in the argument.

10. What does Mike's statement (j) do in the discussion?

- ☐ a. Uses a value to support his position.
- ☐ b. Qualifies or limits his position in the discussion.
- ☐ c. Suggests that Joe's position in the argument needs qualifying or limiting.
- ☐ d. Suggests that a source is unreliable or untrustworthy.
- ☐ e. Suggests that Joe needs more proof for his last statement.

11. What does Mike's statement (k) do in the discussion?

- ☐ a. Suggests that Joe's claim requires evidence and should be tested.
- ☐ b. Shows that Joe's statement is not true.
- ☐ c. Suggests that a qualification already given is not adequate.
- ☐ d. Qualifies or limits a fact or claim.
- ☐ e. Takes a more general position in the discussion.

12. What does Joe's statement (l) do in the discussion?

- ☐ a. States that an important value is being violated.
- ☐ b. Clarifies a word or phrase in the argument.
- ☐ c. Gives a source to support the reliability of a fact or claim.
- ☐ d. Suggests that Mike holds two contradictory values.
- ☐ e. Clarifies his position in the argument.

GO RIGHT ON.

109 eighth grade Concord students is 0.75. This is a split-half estimate corrected by the Brown-Spearman formula. It should be noted that the test was administered after the students had been subjected to an experimental curriculum. Since the test contained only twenty (20) items, presumably it would be practical to increase the number of items substantially and thereby bring the reliability up to a point where individual scores could be meaningfully interpreted. As with all the instruments described here, we treated SIAT No. 1 only as an experimental research tool, and were not concerned with refining it to the point needed for individual diagnosis until we had some indication of whether or not it was sensitive to the effects of systematic instruction.

SIAT No. 2: Argument Description and Rebuttal

Like SIAT No. 1, SIAT No. 2 is a group administered pencil-and-paper test with prestructured multiple choice responses. The purpose of the test, however, is to assess how well the student can describe the substance of an argumentative dialogue (not the intellectual operations which describe it), and whether or not the student can select the best rebuttal statements which might be used to counter statements made in the dialogue. The subtest on labor which follows illustrates the general form of the test. Part A requires the student to select the statement which best summarizes the argument. The purpose is to see to what extent the student can provide himself with a focus of general relevance. Part B requires the student to evaluate summaries of the argument presented in dilemma form. In these statements the argument is raised to a general value level, with two values opposing one another. Again the student is asked to judge which values are most relevant to the argument, and which can most appropriately be used to construe the argument. Part C presumably measures the student's ability to comprehend the substance of the argument at a very specific level. Part D is similar to Part C, except that the student must make a judgment of relevance regarding new information and determine whether it is relevant to one side or the other or to both sides in the argument.

Part E accounts for about one half of the total test and is by far the most important section. To explain this part more fully, let us look at item 13. In evaluating the various responses, the student can analyze the argument by asking himself two questions:

"What type of problem is the argument focused on at just this point in the discussion?" and

"What problem analysis strategy is most appropriate to clarify or resolve this point of disagreement?"

Such an analysis might proceed as follows: Statement (a) asks for specific evidence to support a generalization, and possibly a qualification of what is probably an overgeneralization. From our point of view, the statement is relevant because the generalization being questioned is an important factor in the argument. Statement (b) accuses the antagonist of a position he has not taken. Moreover, the position of which John is accused is stated in

TABLE 10.1

RESPONSES OF HARVARD FIRST YEAR GRADUATE STUDENTS IN SOCIAL STUDIES
EDUCATION ON EXPERIMENTAL SIAT No. 1

Topic of Argument: Space RaceItems, Choices, and Frequencies

1.	2.	3.	4.
a.	a.	*a. 13	a.
b.	*b. 23	b.	b. 4
c.	c.	c. 2	c. 1
*d. 24	d.	d. 2	*d. 17
e.	e. 2	e. 6	e. 2

Topic of Argument: LaborItems, Choices, and Frequencies

1.	2.	3.	4.
*a. 23	a.	a. 5	*a. 21
b.	b.	b. 5	b.
c.	*c. 17	*c. 14	c. 1
d. 1	d. 1	d.	d. 1
e.	e. 6	e.	e. 1

Topic of Argument: Communists
in AmericaItems, Choices, and Frequencies

1.	2.	3.	4.
a. 2	a.	a. 5	a.
b. 2	b.	b. 1	b.
*c. 20	c. 1	c.	c. 2
d.	*d. 23	d.	*d. 22
e.	e.	*e. 20	e.

Topic of Argument: Government
Regulation
of DrugsItems, Choices, and Frequencies

1.	2.	3.	4.
a. 2	a. 1	a. 1	*a. 24
b.	*b. 14	*b. 20	b.
c.	c. 5	c. 3	c.
d. 1	d. 1	d.	d.
*e. 21	e. 3	e.	e.

Topic of Argument: The Algerian
RebelsItems, Choices, and Frequencies

1.	2.	3.	4.
a. 1	*a. 20	a.	a.
*b. 23	b.	b.	b. 1
c.	c.	c.	*c. 21
d.	d.	d.	d. 2
e.	e. 4	*e. 24	e.

*Answers considered correct
by authors

Read the following conversation carefully. You will be asked several questions based on what you read. While answering the questions you may look back as often as necessary.

JOHN AND DICK DISCUSS LABOR

John: Have you read what's happening to the working man these days? He's being taken in by labor unions to the tune of millions of dollars, and all he gets out of it is headaches and strikes.

Dick: So you're an enemy of the working man. Just because the unions have found a way to give the working man a fair break, fellows like you are down on them.

John: I'm no enemy of the working man. I know labor has had a long hard road to travel. And once the unions really tried to help him. But that was years ago. Now the unions are led by a group of sharp gangsters who are only after the worker's pocketbook.

Dick: Any labor expert will tell you that if the unions get weak, business men will grind the working man right down into the dust.

John: Let's face it. The working man is better off than he has ever been. He doesn't need a union leader to stand over his old enemy with a loaded gun when the enemy has been dead for 25 years. What the worker needs today is a law to take the gun away from the gangsters and union leaders who are robbing him.

Dick: Now the truth is out. You want to have Congress outlaw any kind of union that can deal effectively with business. That's just like you, always turning to the government when you want something. We all know that the working man is a long way from getting a fair break, and the union is the only way he can get it.

John: You're a fine one to talk about asking the government for help. You weren't against passing laws to give the worker higher wages and shorter hours and an old age income twenty years ago. I say when unions are packed with gangsters and thugs, it's about time Congress stepped in and put an end to them.

Dick: Maybe what you say is true of a few unions, but can you name any really important unions you think should be outlawed?

John: Sure. The Teamsters, the Longshoremen, and the Steelworkers, to name only a few. You can read all about these rotten unions in any big city newspaper.

(Go on to next page)

- Dick: I knew you would get around to the Teamsters sooner or later. You've had it in for them ever since you had an argument with that truckdriver last year. What about all the honest unions, that everyone knows are honest.
- John: Finding an honest union is like trying to find the corners on a baseball.
- Dick: What about the Rubber Workers and the Auto Workers and lots and lots of others?
- John: Some unions are just able to cover up their dishonesty better than others. We've put up with these blackmailing gangs long enough. The sooner laws are passed to put crooked union leaders out of business, the better off we'll all be.
- Dick: There you go again, off the deep end as usual. You want to throw the baby out with the bath water. Just because a few labor leaders have not always been completely honest doesn't mean we should take away the rights of millions of workers to organize and belong to unions.
- John: The only ones really who want to "organize," as you call it, are a few selfish union leaders who don't want to give up a good thing. You say I want to take away the worker's rights. That's where you're wrong, Dick. I really want to give millions of workers back their rights, not take them away.
- Dick: Yah! You want to give them back their rights by outlawing the unions they have freely joined and supported.

* * *

Part A. Argument Summary

On your answer sheet check the question below which best describes what the argument is about.

1. a. Are modern labor unions useful to business, industry and the public?
- b. Are modern labor unions with their high costs and corrupt leaders really harmful to the worker?
- c. Do workers have a good reason to be against unions and their leaders?

(Go on to next page)

- d. Can dishonest labor leaders be removed from the labor movement without destroying the unions themselves?
- e. Have modern labor unions harmed workers to the point where the government should step in and protect the workers?

Part B. Ideas of Right and Wrong

John and Dick disagree about some important ideas of right and wrong. On your answer sheet check the statement below which best describes their disagreement over what is right and wrong.

- 2. a. Should the government interfere and control unions in the interest of the workers or allow unions to have freedom with the risk of dishonest unions?
- b. Should we allow a little violence and dishonesty in the unions and have a union that can get things for its workers, or have a democratic union that does little for the worker?
- c. Is it better to leave unions alone even though they aren't doing all they can for their members than to attempt to change them and risk the possibility of their being destroyed?
- d. Is it better for dishonest union leaders to get some benefits for their members now than it is for them to "go straight" and hope for a good deal in the future?
- e. Should unions be strong and dictatorial to get high wages for their members or be honest in order not to risk government interference and control?

Part C. Who Said What?

Items 3 through 7 below describe in different words something John said in the argument, something Dick said in the argument, or something that neither or both might have said in the argument. On your answer sheet check D if you think Dick made the statement, check J if you think John made the statement. If you think neither or both might have made the statement, check Can't tell.

- 3. The worker gets more problems than benefits by belonging to a labor union.

(Go on to next page)

4. Without the union the worker risks getting poorer wages and working conditions.
5. The purpose of the union from the very beginning was to get as much as it could from both the workers and the businessmen.
6. The government should help the working man.
7. Labor unions should be more severely controlled.

Part D. Supporting Statements

Items 8 through 12 are statements of fact which you can assume are true. If these statements had been made at any time during the argument, do you think they would have supported Dick's position, John's position, or the position of neither or both. On your answer sheet check D if you think the statement supports Dick's position, check J if you think the statement supports John's position. If you think the statement supports neither or both positions, check Can't tell.

8. Workers who belong to powerful unions are generally paid better than workers who belong to no union at all.
9. Under the leadership of Senator McClellan, the Senate Rackets Committee discovered in 1958-1959 that several large unions had a number of ex-convicts employed as high union officers.
10. Many steelworkers feel that the gains which come from union bargaining with the steel companies are more than lost by having strikes every two or three years.
11. Many textile companies in New England moved South because they didn't want to hire union labor.
12. In 1947 Congress passed the Taft-Hartley Act which gives the President the right to halt a strike for eighty days if it threatens the health and security of the nation.

(Go on to next page)

Part E. Argument Reply

Items 13 through 17 contain statements made by Dick or John in the argument. In this part of the test you are to check the two best replies which you might have made to each statement if you had been in the argument. The best replies are those which may clarify the disagreement or move the argument forward toward some agreement. Remember, for items 13 through 17, check the two best ways to answer each statement.

13. Now the unions are led by a group of sharp gangsters who are after the worker's pocketbook.
- a. Do you think all unions are led by dishonest leaders or do you have some particular unions in mind?
 - b. You're defining gangster as anyone who wants to give the underdog a fair chance.
 - c. Would you clarify exactly what you mean by the worker's pocketbook?
 - d. What evidence do you have that the worker's pocketbook is at all affected by how much money the companies make?
 - e. Do you mean that union leaders are actually breaking laws?
14. Any labor expert will tell you that if the unions get weak businessmen will grind the working man right down into the dust.
- a. That depends on what ideas you have in mind when you talk about labor unions.
 - b. Let's set up a test and find out whether the workers belong to weak or strong unions.
 - c. The evidence shows that businessmen are too reasonable to grind anyone down into the dust.
 - d. Let's set up a test and find out whether the workers in weak unions get lower wages and poorer working conditions than workers in strong unions.
 - e. Who are these labor experts?

(Go on to next page)

15. He (the worker) doesn't need a union leader to stand over his old enemy with a loaded gun when the enemy has been dead for 25 years.
 - a. The unions don't have any weapon as powerful as a loaded gun.
 - b. The owners and managers of companies may be "enemies," to use your expression, but in what way are they "dead?"
 - c. Union leaders had to use guns because they were treated like criminals, and in many places they still are.
 - d. You're making an unfair comparison when you say the union leader is like the man with a loaded gun.
 - e. Calling union leaders names isn't going to get us anywhere.
16. You can read all about these rotten unions in any big city newspaper.
 - a. I guess I read the wrong newspaper.
 - b. Let's look up some of these statements made in these newspapers.
 - c. Isn't there evidence that these big newspapers are just trying to sell the papers and that's the reason they make these charges against unions?
 - d. What do these newspaper reports say?
 - e. Let's define clearly the difference between a big city newspaper and other good newspapers.
17. Just because a few labor leaders haven't always been completely honest doesn't mean we should take away the rights of millions of workers to organize and belong to unions.
 - a. I don't know what other conclusion we could come to considering all the facts we have about modern labor unions.
 - b. You and I have a different understanding of the word "few."
 - c. If workers no longer can protect their own rights, then the government must step in and protect the workers.
 - d. When you say that union leaders haven't always been completely honest, you're practically admitting that unions have to have greater control.
 - e. You're going to have to show me that having the government step in means that workers will lose their rights.

vague and loaded language. Statement (c) is an important intellectual operation in argumentation, but inappropriate at this point in the argument. There is apparently no controversy over the definition of "pocketbook." Statement (d) is also an important intellectual operation, but like (c) is irrelevant at this point in the discussion. No one is arguing over the extent to which the worker's pocketbook is affected by company earnings. Statement (e) asks for a clarification or definition of the term "gangster," which Dick suggests is being used loosely in the argument. It asks an important question and is relevant at this point in the argument. Thus, the most reasonable answers would seem to be (a) and (e).

The comprehensive nature of SIAT No. 2 tends to make it difficult to interpret. Seven items are based mainly on reading comprehension. Five items are based on the student's ability to identify the relevance of various factual statements to the two positions in the argument. Ten points (five items with two points each) are of the type illustrated above. They presumably get at the student's ability to analyze the argument explicitly or intuitively and determine whether rebuttal statements have immediate relevance to the disagreement at hand. It was our assumption (as yet untested) that the three types of abilities required to do the whole test were in ascending level of complexity and difficulty. Thus, a high score could be achieved only if the student were competent at all three levels.

We must confess, however, that the SIAT No. 2 was developed early in our thinking about evaluation. We have since become much more skeptical about the diagnostic or research value of comprehensive instruments. It is probably sounder evaluation strategy to develop sharper assessment instruments which point at very specific abilities. SIAT No. 1, already described, was developed later and illustrates this sharper focus.

We have a number of reliability estimates for the SIAT No. 2. Four different forms were developed, each based on a different topic. The topics were labor leaders (L), racial integration (I), atomic fallout (F), and the quality of American education (E). Since the four forms were administered to Concord students in the fall of 1959 before experimental treatment and in the spring of 1961, after having been exposed to the experimental curriculum, we can correlate the individual forms with each other at the two different dates of administration, as well as correlate the forms with one another. We can, of course, correlate composite scores on two or more forms with one another. Table 10.2 summarizes these comparisons. In general, the reliabilities are relatively low. The labor form administered in the fall of 1959, for example, correlates only 0.26, 0.35, and 0.43 with the three other forms administered at approximately the same time. Increasing the length of the test does, however, increase its reliability substantially. The composite form "Labor and Integration" administered in the fall of 1959 correlates 0.53 with the composite form "Fallout and Education" administered at the same time. Moreover, when all four forms are combined into a single test and correlations estimated for the two different dates, the correlation goes up to 0.60. Considering the fact that almost two years elapsed between the two administrations, this relationship is quite high.

TABLE 10.2

CORRELATIONS OF FORMS OF SIAT No. 2
ADMINISTERED IN NOVEMBER 1959 AND JUNE 1961
TO 109 JUNIOR HIGH SCHOOL STUDENTS

Form	Date	1	2	3	4	5	6	7	8	9	10	11	12	13
1. Lab	N.'59													
2. Int	N.'59	.25												
3. Lab & Int	N.'59	.75	.78											
4. Lab	J.'61	.52	.23	.43										
5. Int	J.'61	.43	.14	.28	.47									
6. Lab & Int	J.'61	.52	.21	.39	.85	.79								
7. Fall	N.'59	.35	.32	.36	.49	.34	.44							
8. Educ	N.'59	.44	.33	.47	.47	.35	.41	.52						
9. Fall & Educ	N.'59	.45	.41	.53	.53	.39	.47	.88	.83					
10. Fall	J.'61	.47	.24	.41	.48	.49	.50	.43	.44	.45				
11. Educ	J.'61	.23	.07	.16	.37	.25	.33	.27	.23	.25	.35			
12. Fall & Educ	J.'61	.44	.2	.35	.53	.47	.52	.44	.42	.43	.86	.78		
13. All forms	N.'59	.71	.59	.83	.53	.39	.48	.57	.73	.82	.57	.28	.53	
14. All forms	J.'61	.58	.25	.45	.78	.73	.85	.51	.48	.53	.78	.63	.87	.6

Topics of Dialogues on which Forms are based:

Lab = Labor

Int = Racial integration

Fall = atomic fallout

Educ = quality of American education

N = November

J = June

SIAT No. 3, Argument Analysis Test

The two SIAT measures described above begin at a common point. They ask the student to read and analyze an argument. SIAT No. 1 presents the student with specific statements in dialogue form and asks him to choose from alternative responses the analytic operation illustrated by a particular statement, or the operation which the response may require the antagonist to perform. The major part of SIAT No. 2 requires the student to evaluate alternative rebuttal responses to statements made in an argumentative dialogue. Both of these tests have a number of the same difficulties which we discussed at length in the section on critical thinking tests in general. First, the student responds to a fragmented and artificial argument. Second, in any objective pencil-and-paper test a great deal of analytical prestructuring takes place before the student is asked to make a response. For example, from the total argument, significant statements are lifted for more careful analysis. The student is therefore not required to make this initial judgment of relevance. Moreover, the student is faced with an array of responses which immediately defines for him the limits within which acceptable or valued answers can be made. Prestructured pencil-and-paper tests can easily be considered teaching as well as measuring devices.

The objective pencil-and-paper format, in short, seems to provide the student with a structure and a set for approaching problems in a reflective and systematic way. The test, therefore, tends to discriminate only whether or not the student knows the analytic operation appropriate to deal with a given situation. It does not, we suspect, measure whether or not the student is disposed to use these operations in less structured, though relevant, situations. In a test in which the student is asked, for example, to estimate appropriate levels of inference, or whether or not a particular conclusion follows from a given statement, the student is specifically told what the problem of analysis is, and moreover, an array of situations is provided which structures the range of examples within which the discrimination is to be made. When the task is structured to this extent and the student undertakes a number of similar tasks (items) in close succession, we feel that general reasoning ability, rather than more specific critical thinking skills, may well play the predominant part in determining how well the student will do. The two SIATs thus far discussed, however, at least have the virtue of requiring the student to consider several intellectual operations at the same time. SIAT No. 2 moreover, has the additional advantage of not telling the student the analytic concepts on which the dialogue analysis should be based. Only the substantive argument and substantive rebuttal statements are presented. It should also be noted that the content upon which the analysis is based is selected from questions of public policy. The fact remains, however, that selecting from a limited number of prestructured choices based on a one page dialogue is quite a different situation from freely analyzing an argument orally or participating in a live argument in which the only structure for analysis and dealing with opposing positions is that provided by the participants.

In an effort to meet these problems presented by the precoded pencil-and-paper test we moved to a relatively unstructured test situation, relying upon the use of systematic content analysis as a method of quantifying student responses. At this point we had to choose between obtaining written protocols by students for later analysis or obtaining tape recorded protocols of oral behavior. The latter course of action was selected for two reasons: (1) our experimentation and training procedures are directed mainly at shaping oral behavior; and (2) it is probably safe to assume that the act of writing itself, as a means of expression, inhibits the more spontaneous elements of the student's responses and adds new elements to the problem of communication. There are, of course, aspects of face-to-face oral interaction that may also inhibit expression, but we assume that this was, at least, a "first order skill" practiced since shortly after infancy.

Two new SIATs were developed in an effort to de-structure the situation in which the test was given. The first was comparable to SIAT No. 1 in the type of intellectual competence assessed. It required the student to differentiate and identify various types of disagreements as well as suggest how these disagreements might be clarified or resolved through different analytic strategies. The test was administered in an individual interview situation. The following directions are given to the student:

We are going to read an imaginary conversation between two men, Bob and Don. Then I am going to ask you to describe the main points of disagreement between the two men, and I am going to ask you how you would go about settling or clarifying these points of disagreement. In analyzing these main points of disagreement try and use any critical thinking skills or knowledge you may have learned about the analysis of controversial issues.

The interviewer then reads the dialogue aloud with the student following on a copy of his own. The student is allowed to refer to his own copy of the dialogue throughout the interview if he desires. Following is the dialogue used in one experimental version of the test.

Working for Mr. Spindle

- 1) Don: The trouble with Mr. Spindle, the manager of our plant, is that he just won't sit down and talk about wages and working conditions with the workers; he doesn't believe in collective bargaining.
- 2) Bob: Of course he does. Every Thursday afternoon he calls a few of the workers together and talks over conditions in the plant.
- 3) Don: That's not collective bargaining. Collective bargaining is when the boss sits down and talks to the official representatives of all the workers, not just a few men.

- 4) Bob: I say collective bargaining is when the boss talks to a group of men about their working conditions. It's as simple as that.
- 5) Don: Well, I don't care about what you call collective bargaining anyway. The fact of the matter is that old man Spindle doesn't bother to meet with any of the men on a regular basis.
- 6) Bob: He met with the men last Thursday.
- 7) Don: I think that's the first time he has met with the men for the last month.
- 8) Bob: Nonsense. I know he meets with the men every week regularly.
- 9) Don: I don't see how you can say that. I talked to four men on Friday, and they told me that the Thursday meeting was the first of its kind for a month or more.
- 10) Bob: What difference does it make, anyway, whether or not Mr. Spindle meets with the men on a regular basis? He is the boss of the plant. He shouldn't have to discuss all his plans with the men who work for him. If they don't like the way the plant is run, they can quit.
- 11) Don: What chance does one man have trying to convince old man Spindle that his working conditions are dangerous or his pay is too low? The men certainly should have some say about working conditions that affect them.

The dialogue is designed to illustrate the three major types of disagreements of problems described in Part Two. Statements three and four are written as an example of a definitional disagreement over the use of the term "collective bargaining." Statements five through nine are written as an example of a factual argument over how often Mr. Spindle met with his men on a regular basis. It is also possible to construe these statements as a disagreement over the use of the word "regularly." Statements ten and eleven presumably illustrate a value conflict over the rights workers have to representation in determining their wages and working conditions as against Mr. Spindle's right as owner to control these decisions. After the dialogue is read, the student is asked two sets of questions. The first is general and open-ended. The student is allowed to analyze the dialogue without specific prompts. The second set of questions focuses on parts of the dialogue which illustrate specifically the types of disagreement described above.

In other words, the dialogue is structured for the student into the appropriate problem types as a first step in analysis. With this additional information the student is again asked what type of disagreement is illustrated by each phase of the dialogue, and how he would go about clarifying or resolving such a controversy. If the student is unable to identify the type of controversy involved, he is given the answer, and again asked for the appropriate way to deal with it. Various prompts are also given the student by the interviewer as he describes how he would deal with each type of problem. The interview schedule is reproduced below.

INTERVIEW SCHEDULE FOR SIAT No. 3

Part One

I GENERAL

A. WHAT ARE THE MAIN POINTS OF DISAGREEMENT BETWEEN B AND D AND HOW WOULD YOU GO ABOUT SETTLING THESE POINTS OF DISAGREEMENT?

B. ARE THERE ANY OTHER MAIN POINTS OF DISAGREEMENT BETWEEN B AND D OR ANYTHING ELSE YOU WOULD LIKE TO SAY ABOUT SETTLING THESE POINTS OF DISAGREEMENT?

Repeat question B. above until the student says he has nothing else to add.

C. NOW WE ARE GOING TO DISCUSS SOME SPECIFIC POINTS IN THE ARGUMENT MORE FULLY.

YOU MAY HAVE TO REPEAT SOME OF THE THINGS YOU HAVE ALREADY TOLD ME TO ANSWER QUESTIONS ON THESE SPECIFIC POINTS. DON'T HESITATE TO REPEAT OR RESTATE ANYTHING YOU HAVE ALREADY SAID, IF YOU THINK IT IS THE RIGHT ANSWER TO ONE OF THE QUESTIONS I ASK.

Part Two

II DEFINITION PROBLEM

A. WHAT DO YOU THINK IS THE MAIN POINT OF DISAGREEMENT IN STATEMENTS () () () ()

If succeeds (says they are disagreeing over the definition or meaning of the word collective bargaining)
go on to B:

If says disagreement is over the meaning of a word, ask student

WHAT WORD ARE THEY DISAGREEING OVER?

If says, disagreeing over collective bargaining, or disagreeing over whether or not there really was collective bargaining, ask student:

IN WHAT WAY ARE THEY DISAGREEING OVER "COLLECTIVE BARGAINING"?

If succeeds, go on to B.

If fails, say:

IS THERE ANY OTHER WAY YOU WOULD DESCRIBE THE DISAGREEMENT IN THESE STATEMENTS?

If succeeds, go on to B.

If partial success, repeat appropriate steps above.

If fails, say:

ONE OTHER KIND OF DISAGREEMENT B AND D ARE HAVING IN THESE STATEMENTS IS OVER THE DEFINITION OR MEANING OF A WORD. WHAT IS THE WORD?

If succeeds, go on to B. If fails, say: THE WORD IS COLLECTIVE BARGAINING.

- B. (If included in answer to A, omit)
IN THE DISAGREEMENT OVER THE MEANING OF THE WORD "COLLECTIVE BARGAINING" WHAT IS B'S DEFINITION OF THE WORD AND WHAT IS D'S DEFINITION OF THE WORD?

(If only gives either B's definition or D's definition, say): YES, THAT IS B'S DEFINITION, NOW WHAT IS D'S DEFINITION (or vice versa)

- C. YES. BOB THINKS THAT COLLECTIVE BARGAINING MEANS THAT THE BOSS OR OWNER TALKS WITH HIS MEN ABOUT WAGES AND WORKING CONDITIONS IN SMALL OR LARGE GROUPS WHETHER THESE GROUPS OF MEN REPRESENT THE WORKERS OFFICIALLY OR NOT. AND DON THINKS THAT COLLECTIVE BARGAINING OCCURS ONLY WHEN THE BOSS TALKS TO THE OFFICIAL REPRESENTATIVES OF ALL THE WORKERS.

IN THIS DISAGREEMENT OVER THE MEANING OR DEFINITION OF THE WORD COLLECTIVE BARGAINING, HOW WOULD YOU GO ABOUT SETTLING THE DISAGREEMENT?

III FACTUAL PROBLEM

- A. WHAT DO YOU THINK IS THE MAIN POINT OF DISAGREEMENT IN STATEMENTS () () () ()

If student succeeds, go on to B. If in doubt, say: I'M NOT QUITE SURE WHAT YOU MEAN.

If student fails, say:

THE DISAGREEMENT SEEMS TO BE OVER A FACT. WHAT IS EACH MAN CLAIMING TO BE TRUE, AND HOW DO THEY DISAGREE?

(The men are disagreeing over whether Mr. Spindle talked regularly with even a small group of men at his plant. One says they met every Thursday. The other says they meet only every once in a while.)

If student succeeds, go on to B.

If student fails, say:

THE MEN ARE DISAGREEING OVER WHETHER OR NOT MR. SPINDLE REALLY MET REGULARLY WITH A GROUP OF HIS EMPLOYEES. BOB CLAIMS THAT THEY DID MEET REGULARLY WITH MR. SPINDLE. DON CLAIMS THAT THEY MET ONLY EVERY ONCE IN A WHILE. THEY ARE DISAGREEING OVER WHAT CLAIM IS TRUE.

- B. NOW THAT WE KNOW WE HAVE A FACTUAL DISAGREEMENT IN THIS PART OF THE ARGUMENT, HOW WOULD YOU GO ABOUT SETTLING SUCH A DISAGREEMENT? THAT IS, WHAT WOULD YOU DO IN THIS PARTICULAR CASE TO SETTLE THE DISAGREEMENT?
- C. ANY OTHER INFORMATION YOU WOULD LIKE TO HAVE TO SETTLE THIS ARGUMENT?
- D. (ARE THERE ANY OTHER WAYS) HOW MIGHT YOU GO ABOUT MAKING SURE THAT THE INFORMATION YOU GOT WAS TRUE OR RELIABLE?
- E. WHAT ARE SOME OF THE (OTHER) THINGS YOU WOULD TAKE INTO ACCOUNT IN DETERMINING WHETHER YOU HAD ENOUGH RELIABLE INFORMATION?

- F. CAN YOU THINK OF ANY (OTHER) SPECIAL PROBLEMS YOU MIGHT RUN INTO IN GATHERING EVIDENCE YOU NEED?

IV VALUE PROBLEM

- A. WHAT DO YOU THINK IS THE MAIN POINT OF DISAGREEMENT IN STATEMENTS () () () () .

1. If specific statement of value disagreement given, ask:

ARE THERE SOME VALUES OR RIGHTS (THAT IS IDEAS OF GOOD OR RIGHT) SUGGESTED IN THIS DISAGREEMENT WHICH B OR D MIGHT GIVE TO SUPPORT HIS POSITION?

If fails, ask:

ARE THERE ANY VALUES OR RIGHTS D MIGHT GIVE TO SUPPORT HIS POSITION? ARE THERE ANY VALUES OR RIGHTS B MIGHT GIVE TO SUPPORT HIS POSITION?

(If the student has already given one value, omit the appropriate question above)

2. If a general statement of value disagreement given, ask:

HOW WERE YOU ABLE TO FIGURE OUT WHAT PRINCIPLE WAS INVOLVED? AND IN WHAT WAY IS (state second value, if given) INVOLVED IN THIS DISAGREEMENT?

- B. Whether or not the student succeeds in part A, summarize the correct answer.

YES. WE SEE THAT THERE IS A VALUE QUESTION AT THIS POINT IN THE ARGUMENT. THE QUESTION IS SHOULD MR. SPINDLE BARGAIN COLLECTIVELY WITH HIS WORKERS, AND GIVE THEM EQUAL BARGAINING POWER, WHEN IN FACT HE OWNS AND RUNS HIS OWN FACTORY. TWO IMPORTANT VALUES INVOLVED HERE MIGHT BE THE RIGHT TO CONTROL THE PROPERTY YOU OWN AND USE IT AS YOU SEE FIT, SINCE MR. SPINDLE DOES OWN THE FACTORY: AND THE RIGHT OF THE WORKERS TO HAVE AN EQUAL SAY IN DETERMINING THEIR WAGES AND WORKING CONDITIONS, SINCE THEY ARE DEPENDENT ON THEIR JOBS FOR A LIVING.

NOW, WHAT ARE SOME OF THE THINGS WE MIGHT DO TO SETTLE THIS DISAGREEMENT OVER VALUES?

- C. ARE THERE ANY OTHER THINGS WE MIGHT DO?

The interview procedure has an obvious advantage over precoded pencil-and-paper tests in that the student is not necessarily penalized for seeing different problems from those which the authors of the test have deliberately planted in the dialogue. The scorer of a multiple choice test can have no knowledge of what the student is thinking as he takes the test. This is a special problem in the interpretation of a complex message or dialogue. To assume that the author's interpretation of the problem presented by statements in the dialogue is the only correct interpretation is out of line with our common knowledge.

SIAT No. 3 was scored on the basis of three different types of responses:

(1) Judgment was made concerning whether or not the student was able to identify the various types of disagreements illustrated by the dialogue. The degree of explicitness with which the disagreements were identified was taken into account. If the student said, for example, that statements three and four constitute a definitional argument over the meaning of the term "collective bargaining" he was given full credit. If the student said that the argument was over what "collective bargaining" is, he was given partial credit, since we considered "is" less explicit than "meaning of" or "definition of." If the student described the definition only in terms of the specific situation, he was given less credit, e.g., if the student said that in statements five through nine there was a controversy over whether or not Mr. Spindle held regular meetings with his men.

In identifying the value disagreement in statements ten and eleven, the student was given additional credit if he not only identified the type of disagreement, but indicated what values were involved in the disagreement. (The values suggested as a basis for the controversy had to be stated in a plausible context, i.e., the scorer had to feel that the student understood why these values were central to the conflict; he could not simply list value labels he thought might be relevant.)

(2) We evaluated the student's ability to deal with the controversy in a constructive fashion quantitatively by counting the number of "reasonable" strategies presented. While we were able to anticipate most of these strategies in advance of the interviewing, post hoc categories were added during the scoring. Wide latitude was allowed in scoring this part of the interview; the student was given credit for any reasonable response whether it was substantive or analytic. This is particularly important for evaluating experimental teaching procedures in which students may be taught specific patterns of analysis. Control students must not be penalized for using "common sense responses" when these responses, in the opinion of the scorer would, in fact, clarify or move the argument along. Scorers for this test were obtained from outside the Project and did not know which were the experimental students.

(3) A number of important analytic concepts were selected as particularly important for dealing with the controversy. If the student used these concepts either explicitly or implicitly he was given additional credit. These concepts included hypothesis, evidence, assumption, etc. (The complete list is given on page one of the interview scoring sheet below.)

In addition to basing scores on these three general areas of learning (problem or disagreement identification, identification of strategies for resolving disagreement, and use of analytic concepts), the student could get more or less credit depending on the point in the interview that a valued response was given. A response stated after the first open-ended question received more credit than a response given after the dialogue had been structured into its three major parts. Moreover, after the student had approached the more specific aspects of the problem in the second part of the interview, he was given more credit if fewer prompts were required to elicit a valued response. For example, in the analysis of the factual problem in the dialogue, an initial question and four prompts were given when the student was asked about strategies for resolving or clarifying the problem. More credit was received if a valued response was given after the first prompt than after the fourth prompt. The reader can perhaps get a better idea of the scoring procedure by looking at the first page of the scoring booklet. This is filled out after the first general open-ended question. Additional scoring sheets are required to record the number of prompts given before the student makes a scorable response.

We had little trouble obtaining reliability in the scoring of this measure. Pairs of scorers were trained until they consistently reached 85% agreement on the total number of items scored. This required from 8 to 12 hours of training. The only reliability estimate we have for SIAT No. 3 is between an experimental pretest of the instrument administered in the fall of 1960, and the regular test administered in the spring of 1961. The correlation is a discouraging 0.35. However, the experimental pretest was given precisely to work the difficulties out of such a measure, and the administration lacked adequate standardization. The hypothesis that the low correlation is largely due to the unreliability of the experimental pretest rather than the unreliability of the final form of SIAT No. 3 itself is further supported by the correlation between SIAT No. 1 (a similar pencil-and-paper test), the experimental version of No. 3 and the final version of SIAT No. 3. These data are presented in Table 10.3. The experimental version also consistently correlates very low with a large number of other measures requiring general reasoning ability, while SIAT No. 3 consistently correlates in the .50s. This is further ground for believing that SIAT No. 3, while admittedly a crude measure in its present form, has the potential of more satisfactory reliability.

SIAT No. 4, A System for Analyzing Free Discussion

SIAT No. 3, although it contains less structure and fewer built-in training procedures than comparable pencil-and-paper measures, still has an essential limitation. It requires the student to deal with an issue only at an analytical level. He is not asked to take a position himself, to defend this position, and to rebut statements attacking his position or supporting other positions. From our point of view, the most natural situation within which to place the student in order to evaluate his analytic and persuasive competence is one involving free oral argumentation. This kind of argumentation must, however, be evaluated by general subjective ratings or by systematic

SCORING SHEET FOR SIAT No. 3

NAME _____

SCHOOL _____

OPEN END

GENERAL CONCEPT

Main Points of Disagreementexplicit implicit specific situationDefinition: Collective Bargaining
or "regularly"Factual: Did Mr. S. meet men
regularly?Value: Should Mr. S. meet
men regularly?

_____	_____	_____
_____	_____	_____
_____	_____	_____

Identifies values: e.g., general welfare _____ equality _____ property/contract _____
other (specify) _____StrategiesDefinition

- _____ find out what collective bargaining means
- _____ identify points of disagreement
- _____ have each elaborate his definition
- _____ see how others use word
- _____ get authority
- _____ dictionary
- _____ contract
- _____ law
- _____ "expert"
- _____ look at other plants
- _____ get men to agree on one meaning
- _____ definition a matter of opinion
- _____ poll men or vote on meaning
- _____ sees problems in own position
- _____ procedural

Factual

- _____ refer to evidence in case
- _____ ask men
- _____ ask Spindle
- _____ question evidence in case
- _____ suggest get more evidence
- _____ check records
- _____ just find out
- _____ procedural

Value

- _____ prove one value's validity
- _____ test factual assumptions
- _____ work toward compromise
- _____ decide which is more important value
- _____ load or exaggerate a value
- _____ test consequences of each value
- _____ define values more specifically in terms of policy decisions
- _____ procedural

ConceptsDefinition (only if not evidence of above)

- _____ criteria or criterial definition
- _____ labeling or pointing definition
- _____ loaded definition
- _____ classification

Factual

- _____ claim
- _____ specific claim
- _____ general claim
- _____ generalization
- _____ tentativity of claims: levels of probability
- _____ hypothesis
- _____ assumption
- _____ testing a claim
- _____ sampling
- _____ representative sampling
- _____ contradiction or consistency
- _____ explanation
- _____ evidence
- _____ intuition
- _____ authority
- _____ observation

Value

- _____ value judgment
- _____ value labels: general welfare _____
- _____ property/contract _____
- _____ equality _____ other (specify) _____
- _____ analogy
- _____ dilemma

(Note: list other responses which seem reasonable.)

TABLE 10.3

CORRELATIONS BETWEEN THREE FORMS OF SIAT

	1.	2.
1. SIAT No. 1 (pencil-and-paper)		
2. SIAT No. 3 (interview)	.60	
3. Experimental pretest of SIAT No. 3 (interview)	.25	.35

content analysis of the interaction process. We have chosen to explore the latter approach in our own work. Following a description of the types of discussion setting in which we obtained student behavior, we shall present in some detail a description of the content analytical system developed to assess that behavior.

Types of Discussion Situations Subjected to Content Analysis

The system for content analysis discussed below is used to score tape recorded rather than live discussions, although with some simplification it could be used for analysis of live situations. It has been used to score three different types of discussion, all based on a controversial case: interviewer-student discussions; discussions composed only of students; and teacher-led instructional discussions. Both the interview situation and student-led discussions were used for evaluation. Teacher-led instructional discussions were taped to check the long term consistency with which teachers can play different teaching roles, a problem which will be discussed in Chapter Twelve. Here we will be concerned with the system as it was used for evaluation.

Student-led discussions. Our initial thoughts about what would be the most appropriate test situation in which to evaluate the student's ability to analyze a controversial case brought us to the student-led discussion. This decision was based primarily on the assumption that the intellectual leadership of the adult teacher and the effects of approving or disapproving teacher cues would be minimized in a discussion group composed entirely of students. In such a situation approximately twelve students are seated in a circle or semi-circle. Each student is given a copy of the controversial case; the case is read, and the students are then asked to arrive at a consensus regarding what is to be done regarding the problem in the case within a stated period of time. Consensus is an important requirement of the task to prevent the students from "solving" the issue by simply taking a vote, and letting the majority rule. (This procedure was adopted from Bales' work with five-man groups.) In general, the discussion that ensues is witnessed by no one except members of the group. The teacher or experimenter reads the initial directions and leaves the room. He enters again only at the end of the period allowed for the discussion to hear the decision at which the group has arrived. (The discussion is, of course, recorded.)

The Socratic interview. The student discussion group as an object of evaluation has certain obvious difficulties: (1) Different groups may have different degrees of interest in the task assigned and may experience procedural problems of different degrees of intensity; (2) If we choose to treat the group as the unit of analysis, it is impossible to tell which and how many members of the group facilitate or inhibit the clarification or resolution of the problem under discussion; and (3) If we choose to evaluate individual responses in the group situation, we face the problem of equating the responses of high and low participants, or even ensuring that some students will respond at all.

These problems can be largely overcome by setting up a situation in which a single adult discusses a controversial case with an individual student. This is essentially a two man group in which we attempt to control the level of sophistication of the person in the group who is not being evaluated. As we have used this situation each interviewer is given the freedom to pursue whatever issues the student chooses to raise. Each interviewer is, however, provided with a "brief" setting forth major issues and arguments as well as critical analogies with which to confront the student. On this and the next page is presented a case with which we have had considerable experience.

A LAW ON HOUSING RIGHTS

Many states have become concerned about the problem of protecting the rights of minority groups, especially the rights of Negroes. Below are excerpts from an imaginary law similar to laws recently proposed in some states. Following the description of the law are comments from two newspaper editorials, discussing the merits and shortcomings of the proposed law.

After we read this information, I want to discuss your opinion of the law, and how you might defend your opinion. I would encourage you to use any discussion skills or critical thinking skills which might help you think through this issue more intelligently.

* * *

A law has been proposed in the state legislature providing that "No person shall refuse to sell, rent, lease, or sublet a house or apartment to any person because of race or color." The proposed law further states that it would be "enforced immediately, except when an owner or landlord can show that undue hardship will result from such immediate enforcement. In cases where undue hardship can be demonstrated, a reasonable delay may be granted."

An editorial in the "Southern Evening Gazette" commented on the proposed law as follows:

This law is based on the false assumption that there is no difference between the two races. The fact is that police records show that, in relation to the population, more Negroes commit crimes of violence than do white people. We also know that broken homes are much more common among non-whites than among white people, and many more Negro children have parents who have neglected or deserted them. Especially important to the question of renting or selling houses is the fact that the yearly income of the average Negro family is much less than that of white families. Also, a larger percentage of Negroes

is out of work each year. With these facts in mind, let us conclude by saying that no matter how much our "liberal" friends do not like to admit it, Negroes are different, and whites have the right to act on this basis. This proposed law takes away that right.

An editorial in the "Evening Sun" took another position:

There are good reasons to support this law. A recent survey in a large city showed that many Negroes with good jobs and a good income want to move to the less crowded suburbs, but are refused housing by white owners. We know of instances in which nationally known Negro artists and athletes have been denied housing in white areas. This is a national disgrace. Sociologists support the position that Negroes often fail to live up to high moral standards because they accept the same opinion of themselves as whites show toward them. Thus, when whites continue to treat them as inferior human beings, Negroes accept the same image of themselves and fail to live up to their full potential. Furthermore, it is common knowledge that Negroes indulge in luxury items such as Cadillacs, fur coats, and expensive clothes, because they are unable to buy or rent decent housing which many want and can afford. We must conclude that this law is needed to right the wrongs committed against the colored race for the last 100 years.

* * *

DO YOU THINK THIS LAW SHOULD BE PASSED?

Having described the contexts within which evaluation has been attempted by the Project, we shall now go on to describe the content analysis system used to quantify the behavior elicited in these various situations.

Quantifying Selected Conceptual Operations Required to Clarify and Defend a Controversial Position

The system we are about to describe is set up to identify a number of the major analytic operations or concepts described earlier in this chapter. Those that are particularly appropriate for the content analytic system are summarized below:

- a. identifying different types of disagreements within a political controversy
- b. using appropriate strategies for dealing with different types of disagreements
- c. using appropriate dialectical operations to explore a disagreement
- d. dealing with the problems of relevance

A general approach to content analysis.⁴⁸ Before actually getting into the substance of the system, we should make some general statements about our approach to content analysis as applied to oral discussion. While some of these statements may seem technical, they are presented for two reasons: They may be helpful to those who wish to experiment with this approach to the evaluation of the student-teacher dialogue, and they will reveal the many problems which must be considered if one wishes to develop such a system. Moreover, the handling of these problems forms the basis upon which one builds a set of assumptions underlying the evaluation of the instrument.

Systematic analysis of interaction, as we use the methodology, involves analyzing ongoing interaction into discrete units which are then categorized. There are three important considerations which must be taken into account in carrying out this process: Into what size units will the total train of interaction be broken? What is the frame of reference of the person who does the categorization? What is the specific nature of the categories used to describe the interaction?

Theoretically, the unit can range in size from an entire meeting or discussion to a particular segment of the discussion or meeting. This segment may be defined in terms of time, a completed verbal interchange, a bit of participation by an individual, or according to some linguistic convention. In general, the unit of our present system is defined by Bales:

The unit to be scored is the smallest discriminable segment of verbal behavior to which the observer, using the present set of categories after appropriate training, can assign a classification under conditions of continuous serial scoring. This unit may be called an act, or more properly, a single interaction, since all acts in the present scheme are regarded as interactions. The unit as defined here has also been called the single item of thought Often the unit will be a single sentence expressing or conveying a complete simple thought.⁴⁹

⁴⁸Our orientation toward content analysis is clearly influenced by Robert F. Bales, Interaction Process Analysis (Cambridge, Mass.: Addison-Wesley Press, 1951). The instrument described here is similar to Bales' in the units of analysis and the observer's frame of reference.

⁴⁹Ibid., p. 37.

We are interested, then, in classifying the "single item of thought." Examples of complete units would be:

"I am sure that the Southerners would not accept immediate integration."

"They are the ones who should do something about the situation."

In general, compound sentences are scored as two units; complex sentences as one. In some instances, several sentences may constitute one unit of thought, e.g., in presenting a single case situation or analogy.

Determining the observer's frame of reference poses a number of questions. One is deciding what the observer's point of view will be toward the group. The observer can

....think of himself as a generalized group member, or, insofar as he can, as the specific other to whom the actor is talking, or toward whom the actor's behavior is directed, or by whom the actor's behavior is perceived. The observer then endeavors to classify the act of the actor according to its instrumental or expressive significance to the other group member.⁵⁰

Another point of view is that described by Steinzor in which the purpose of the observer is to determine the intent of the actor.⁵¹ A third point of view is that of the observer who is to be aloof from the process and not concerned with intent or the effects upon the group or its members. In the observational scheme of Heyns "the observer is outside the process and views each contribution in terms of its theoretical properties as a problem solving function."⁵² Carter and his associates use a scheme in which the observer is not to be concerned with intent or effect, but with the functional significance of an act for the discussion situation through the eyes of an outsider.

In one of the schemes which we have used to quantify teacher style, socio-emotional categories requiring inferences about affective states of mind are included. In using these categories, the observer is to adopt Bales' position in regard to point of view. That is, he is to put himself in the position of the individual toward whom the act is directed and ask himself, "How would I perceive the actor's intentions if I were the recipient of

⁵⁰Ibid., p. 39.

⁵¹B. Steinzor, "The Development and Evaluation of a Measure of Social Interaction," Human Relations, II (1949).

⁵²R. W. Heyns and R. Lippit, "Systematic Observation Techniques," Vol. 1 of Handbook of Social Psychology, ed. by Lindsay Gardner (Cambridge, Mass.: Addison-Wesley Press, 1954).

that act?" or "What would that segment of behavior tell me about the actor's state of mind if I were the one toward whom it was directed?" However, in the present scheme the orientation is different. The purpose of observation is not to make inferences about affective states, but to look for cues which will indicate whether or not the actor is using desired categories of thought. The observer's point of view, therefore, is much like that of Heyns in that the observer serves as an expert in applying criteria of thought categories to the actor's statements. He is "outside the process," except as the discussion context is necessary to apply the criteria. For example, in deciding whether or not a student has stated a qualified value judgment or raised a question of relevance, the observer refers to the content of the statement, not to the manner in which other students might interpret it. As a matter of fact, our experiences confirm Bales' report that for scoring most acts the point of view of the observer is of small importance in obtaining interobserver reliability.

A second aspect of frame of reference is the extent to which the observer should take into account any prior knowledge he has of the group or of the individuals within the group. Our position is that it would be ideal if the observer scored each discussion with no prior knowledge of individuals or group involved. We instruct the observer to try and forget all prior experiences with participants. If an interact arouses a prejudgment the observer is to control it and the question becomes, "How would I score that act if I had never before heard this person?"

A third consideration in defining a frame of reference for the observer concerns the context of the discussion. How much of the context of a particular discussion should the observer take into account in classifying a particular act? Should the act be scored in isolation? Or should the act be scored in its relationship only to the previous act? For example, the statement, "I would agree with John" might be scored disagreement, if John's antagonist in the discussion had just spoken, but scored as "agreement" if the context of our scheme is the total discussion. Since our system uses two scoring systems superimposed on each other, it uses two contexts: one for what we call static categories; the other for dynamic categories. The dynamic system (See Table 10.4) consists of categories which explicitly require the scorer to deal with a context beyond the statement being categorized. (These are essentially dialectical operations.) This context may include one or several other sentences. Scoring in these categories is determined by relationships within or among statements. The static categories (See Table 10.5) theoretically can be scored without taking into account any context beyond the scorable unit. Every unit of behavior is scored in a static category. Dynamic operations are scored only when they are identified. Thus, when a dynamic operation is scored, a double categorization of the same unit occurs.

There are some exceptions, however, to the distinction between static and dynamic categories. The category "relevance," for example, is a dynamic category, but is scored as if it were static because the assertion or

TABLE 10.4

DYNAMIC CATEGORIES

-
1. **CONSISTENCY-INCONSISTENCY:** Statements that indicate explicitly or implicitly that the speaker is aware of a real or possible consistency or inconsistency within his own or another speaker's position. The inconsistency may be between two values, two facts, or two definitions.
 2. **SPECIFICATION and GENERALIZATION:** Specification occurs when the speaker gives a specific statement to illustrate or support a more general statement. Generalization occurs when the speaker draws a more general conclusion from one or more specific statements already given.

Example of specification: "Desegregation is not going well. Only 7% of the Negro children in the South are now going to integrated schools after seven years of illegal segregation." The second sentence would be scored as the static operation "specific claim" and the dynamic operation "specification."

Example of a generalization: "After World War II, Russia 'captured' the countries of eastern Europe, helped China to become a Communist nation, and tried its best to take over Greece and Turkey. Russia is the greatest empirialist nation the world has ever known." Statement two would be scored as a static operation "general claim," and as a dynamic operation "generalization."

3. **QUALIFYING:** A statement which deals with an implicit or explicit inconsistency by pointing out under what general circumstances an exception to a general principle is allowable or possible we score as a qualifying act.

Example: Mr. A: Our civil liberties are our most precious asset. To try and restrict them for any citizen is unAmerican.

Mr. B: If you had been in Germany in the early 1930's, would you have restricted some of the civil liberties granted Hitler when he was conducting mass hate meetings?

Mr. A: I very well might have. I would say that civil liberties should be restricted, however, only when the government which is pledged to protect them is in real danger from an undemocratic and brutal force, which would destroy all civil liberties.

Mr. A's modified position would be scored as static operation "general value judgment," and dynamic operation "qualification."

TABLE 10.5

STATIC CATEGORIES

GENERAL VALUE JUDGMENTS: Statements in which the speaker expresses a preference for a person, object or position in the argument in terms of a general social or legal value, such as: personal privacy, property, contract, speech, religion, general welfare of the group, equality, justice, brotherhood, due process, consent and representation, etc. "Mr. Kohler certainly should have the right to use his property as he sees fit and to make contracts with his workers without union interference."

SPECIFIC VALUE JUDGMENTS: Statements in which the speaker expresses a preference for a person, object or position in the argument in terms of the specific case under discussion. "I think Mr. Kohler should have met the demands of the United Auto Workers."

GENERAL LEGAL CLAIM: Statements in which the speaker asserts that someone has a legal right to do something, expressed in terms of a general legal principle, such as: rule of law, due process, equal protection under the law, constitutional restraints, etc. "He has a right to a fair trial under the United States Constitution."

SPECIFIC LEGAL CLAIM: Statements in which the speaker asserts that someone has a legal right to do something, but does not give a legal principle as a basis for the right. "Mr. Kohler has a right to fire any worker he wants."

GENERAL FACTUAL CLAIMS: Causal, descriptive, or predictive generalizations. "Negroes are just as intelligent as whites."

SPECIFIC FACTUAL CLAIMS: Statements describing specific events delineated in time and space. "The first attempt at integration in Little Rock was on September 4, 1957."

SOURCE: A statement or part of a statement describing the source on which a claim, definition or value judgment is based. "Emergency is defined this way in Webster's New International Dictionary."

DEFINITIONAL CLAIM: A statement about how a word or phrase is defined or should be defined. It is also a statement of analysis by which several meanings of a single word or statement might be distinguished. "An emergency occurs when one or more people are in danger of being injured or losing their lives and property."

TABLE 10.5 -- Continued

REPETITION: A statement in which the speaker repeats himself or communicates something already stated in order to focus the discussion.

CASE: A set of statements which describes specific, real, or hypothetical situations analogous to the one under discussion. Its main purpose is to elaborate the range of situations to which one might apply a value judgment. "Suppose Negroes and whites were given schools of equal quality, teachers of equal quality, books and educational facilities of equal quality: Would Negro schools still be inferior to white schools?"

RELEVANCE: A statement which explicitly deals with the way a statement or group of statements is related to the total argument or to the specific point under discussion. "I don't see what that statement has to do with the discussion."

DEBATE STRATEGY: Ad hominem or other remarks which explicitly discuss the tactics being used by a discussant. "You're just trying to confuse me."

TASK -- PROCEDURAL: A statement directed at controlling the immediate interpersonal situation, and which assumes that everyone in the discussion is trying to do a conscientious job. "Let's take a vote." "Let's give everyone a chance to talk."

DEVIANCE CONTROL -- PROCEDURAL: A statement directed at controlling the immediate interpersonal situation, and assuming that one or more people are violating group norms. "Get back in your seat and sit down." "You don't have to shout."

questioning of relevance usually contains an obvious cue within the statement itself, and because there is often no static category which can be appropriately scored with it.⁵³

Posture of the speaker. "Posture" refers to the attitude of the speaker toward the statement he is making or the function which that statement is performing for the speaker. We have identified and used four postures: declarative statements; interrogative statements; statements which question or express doubt about a prior statement (often in either the declarative or interrogative form, but with an overtone of argumentative intent); and, statements which express self-doubt (as, for example, uncertainty as to the validity of a claim which has been or is going to be made by the speaker.) The posture of the speaker is scored with a symbol within the space provided on a scoring sheet for the appropriate static category.

Orientation of the speaker to the discussion: analysis versus persuasion. We also distinguish and score whether or not the speaker is trying to persuade other group members that his substantive position in the argument is correct, or whether he is attempting to stay "outside" the argument and simply analyze how the group might construe the issues in the case. For example, "That person in the case should not have been allowed to speak because avoiding a riot is more important than his right to speak," is scored as persuasive. The statement, "The problem here is that the principles of freedom of speech and peace and order are both involved in the situation, and we must decide which value should be given greater weight in this instance," is scored as analytical.

Validity and reliability. Initially, of course, using the system to categorize statements in a discussion results in an abstract cognitive description of the discussion. This description must be translated into a quantitative score by determining which categories seem valuable from the point of view of our objectives, and then counting the frequency with which units are scored in these categories. This selection of valued categories is essentially a question of validity. Thus far the system appears to have not only intuitive or face validity, but also reflects the effect of experimental training in reflective thinking. Data on this point will be presented in Chapter Eleven. We have undertaken procedures by which validity can be more firmly

⁵³At this point it should be noted that both the interaction system being discussed here and the one we used to describe teaching styles differ markedly from Bales' in their use of double scoring. That is, each act is scored in at least two subsystems simultaneously. Multiple scoring is possible largely because we score from tapes which allow us to control the rate of scoring. It would be much more difficult to use a system requiring multiple scoring in a live situation. For purposes of analysis, it is important to use an observer scoring sheet which tells us what particular acts have been multiple scored, so that we can distinguish the total units of behavior from the number of categorizations made.

established and preliminary results seem to bear out our faith in the instrument. Presently, we feel that the following categories have value for a discussion involving political controversy.

Static Categories

General Value Judgments and General Legal Claims are valued because they allow the student to deal with the controversial case at a more abstract and general level.

Specific Factual Claims and Sources are valued because they are appropriate ways of supporting more general claims. They are an important part of the empirical proof process.

Definitional Claims are valued because they tend to demand or give greater precision to the various positions in the argument.

Repetition is not valued, since it involves mainly statements which repeat something already said. When the student clarifies by drawing finer distinctions between positions or terms in the argument, it is scored as a Definitional Claim.

Case is valued because, by definition, it is an attempt to expose the point at which an individual will reverse his position, given an array of similar situations to judge. It is essentially a defining operation.

Relevance is valued because it indicates that the student is attempting to deal with the relationship between a particular statement and some larger facet of the total argument.

Dynamic Categories. For obvious reasons, all three dynamic operations are valued. They have been selected for scoring precisely because we think they are important.

Orientation to Discussion. The analytic orientation to the discussion is valued because it tends to indicate that the student is attempting to stand back from the immediate persuasive aspects of the argument and provide a more impartial framework by which to deal with the controversy. The questioning posture may be valued especially in unsophisticated groups when it tends to require discussants to clarify or support a position.

It should be noted that these valued acts are not simply the product of a priori guessing about what acts operate to produce the most intelligent discussion. In arriving at our present position, we have listened to many discussions and done a good deal of cutting and fitting to make our quantitative scoring procedures consistent with our intuitive judgments about what behavior is actually important for clarifying a controversial situation.

Although validity is based mainly on these subjective judgments, we have carried out more systematic work on reliability to establish whether or not the subtleties of language can be objectively scored with these gross categories. Reliability, of course, has two meanings in this context. It can refer to the consistency of behavior under observation or to the consistency with which behavior is observed or categorized. It is the latter with which we are concerned at this point. Initially, as part of the training procedure, agreement among observers in the frequency of units assigned to specific categories was checked by a graphic method.⁵⁴ Having reached an acceptable level of agreement as estimated by this method, we turned to the agreement between observers on the total number of valued acts which should be credited to each student. The degree of association was estimated using the product-moment correlation. Initially four persons were trained to use the system. Each scorer was paired with every other scorer, so that six scoring combinations resulted. The discussions scored were Socratic interviews between a student and adult interviewer, in which the student was challenged to defend a position on a controversial case. The number of discussions scored by each combination ranged from 10 to 18. The results are shown in Table 10.6. There is no widely accepted criterion for the acceptance of such coefficients as satisfactory; as Heyns and Zander⁵⁵ point out, whether one demands a correlation of .70 or .90 is contingent upon the uses to which the observational scores are to be put. As we are now reporting our system within a specific research context, it seems sufficient to point out that with the exception of one coefficient all approach at least .70, with two greater than .80, and one greater than .90. On the average, there is a relatively high level of agreement.

A second reliability study was carried out on a larger sample of discussions, computed by individual valued acts. In this case the scoring was done by two men, one an undergraduate at Harvard College and the other a student at the Harvard Graduate School of Education. Neither scorer knew the purpose of the scoring system or the distinction between valued and non-valued acts. The situation scored was pupil-led discussions in groups of 10-14 students based on controversial cases. The dynamic operations are not included because of the low frequency in these categories. (Between one-third and two-thirds of the discussions contain a frequency of less than three on the three dynamic categories.) These data are presented below in Table 10.7.

⁵⁴Binomial probability paper as developed by Frederick Mosteller and J. W. Tuckey and reported in "The Uses and Usefulness of Binomial Probability Paper," American Statistical Association Journal, XXXIV, 1949, pp. 274-212. For a statement of its application to systematic observation, see R. F. Bates, Interaction Process Analysis, pp. 111-112.

⁵⁵R. W. Heyns and A. F. Zander, "Observation of Group Behavior," Research Methods in the Behavioral Sciences, ed. by L. Festinger and D. Katz (New York: Dryden Press, 1953), p. 411.

TABLE 10.5

RELIABILITY ESTIMATES FOR FOUR OBSERVERS
USING THE CATEGORY SYSTEM

	A	B	C	D
A				
B	.55			
C	.82	.93		
D	.87	.69	.68	

TABLE 10.7

RELIABILITY BETWEEN TWO SCORERS FOR SPECIFIC VALUED ACTS ON THE
SIAT No. 4 FOR 32 STUDENT-LED DISCUSSIONS

<u>Valued Acts</u> <u>r</u>	<u>Correlation (r)</u>
General Value Judgment	.76
General Legal Claim	.79
Specific Claim	.71
Source	.77
General Definition	.86
Specific Definition	.86
Case or Analogy	.73
Relevance	.51
Questioning Posture	.60
Analysis	.42
Total Valued Acts	.89

The reliability of the three low frequency dynamic (dialectical) categories was tested for individual discussions on binomial probability paper. In the 32 discussions scores fell outside acceptable limits three times for generalization-specification, twice for qualification, and three times for consistency-inconsistency.

It should also be reported here that this scoring was done immediately after a training period. We found that scorers tended to become unreliable after a relatively short period of independent scoring, creating some very difficult problems which will be discussed in Chapter Eleven.

We have noted that in applying the concept of reliability to the quantification of oral or written narrative materials through systematic content analysis a distinction must be made between the reliability of the behavior being measured and the reliability with which observers can categorize that behavior. The latter type of reliability provides very little problem with most paper-and-pencil tests because there is little room for error or disagreement once scoring keys have been adopted. The first type of reliability also provides little trouble because it is clear that the behavior referred to is that involved in making responses to the test. That is, the question is, "Can we predict from the responses on one test how the individual, or group of individuals, will respond on the same test at a different time?" This form of reliability is, however, fraught with ambiguities in the systematic observation situation because the observer's interpretations intercede between the individual's behavior and the categorizations which are quantified to obtain a test score. It is often not clear whether in speaking of consistency of behavior we are referring to the actual behavior manifested by the individuals being tested or to the categorizations of that behavior which result from observer behavior. Of course, in either case the reliability of the observer in categorizing behavior will affect the quantifications which we must use in our estimation.

Above we have treated the reliability with which observers can apply the category system. The question we would like to deal with now is whether or not, within the limitations of observer reliability, the observational instrument produces an estimation of behavior which is consistent over time. A factor here, of course, is whether the sample of behavior categorized is large enough to serve as a basis for prediction about future categorizations which will be obtained with the individual or group. With paper and pencil measures this reliability is commonly estimated by correlating individuals' scores derived either from odd versus even items on a test, or from the administration of two forms of a test at two different times. To make an estimate of internal reliability similar to the odd-even comparison, we obtained two scores for each student in the Socratic interview situation. Each scoring sheet, containing spaces for scoring fifteen acts, was numbered in serial order and then used in that order in scoring the interviews. We summed scores on the odd scoring sheets and correlated these with the sums obtained from the even sheets. The correlation obtained is .67, corrected by the Spearman-Brown formula for total valued acts.

Some Concluding Statements on Evaluation

We would conclude this chapter on evaluating competence in political analysis by stating some of the major principles which have developed out of our own work.

1. The criterion of competence in political analysis must be established in a less structured and more realistic setting than that allowed by the multiple choice pencil-and-paper test. Our own evidence indicates that there may be little or no relationship between competence to defend one's point of view in public, and competence required in any one of the common "critical thinking" tests.
2. The use of content analysis to assess competence in political analysis is complicated and fraught with reliability problems. The reliability problems are two-fold: (a) How much behavior does one need to measure before one can make reliable predictions about a person's analytic competence? and (b) To what extent can people agree on how this behavior should be described quantitatively? Nevertheless, we feel these reliability problems can be overcome sufficiently to allow the use of such a system as a criterion index against which to compare less costly and less complex methods of measurement. To proliferate the measurement field with the simple measurement tools already available, which are so heavily saturated with general reasoning and verbal factors makes little contribution.
3. The translation of the objectives of our pedagogical approach to political controversy into specific learning outcomes which can be measured with a set of categories such as described above presents, we believe, unusual possibilities for curricular evaluation. Because learning outcomes can be measured in a situation less structured than paper-and-pencil tests and approaching more closely the circumstances in which the desired concepts will later be applied, the results take on greater meaning and validity. Our reliability data suggest the feasibility of this approach to assessment both in experimentation and classroom teaching. It should be noted, too, that just as a teacher might during any one period of time teach for only one or a few of the concepts included in the category set, so might the set be modified to include fewer categories in order to simplify scoring.

There is, however, no denying the impracticability of careful content analysis for the day to day needs of the average classroom. Teachers, in general, have neither the research competence nor the time to learn and use such a complex system. Ultimately, however, the more complex instrument might be used to establish the validity of simpler category systems, or even of pencil-and-paper tests. There is little doubt in our own minds that present methods of measurement which attempt to assess the process of reflective thinking

with a series of fragmented multiple choice items show insufficient respect for the subtlety and complexity of this competence. It is our conviction that measurement programs will become more significant to teachers and research people when evaluation begins with a recognition of the complexity of the phenomena they are attempting to describe and assess.

4. We should bear in mind that content analysis itself, while it may give us a more reliable picture of an interaction sequence, will not tell us what types of actions or sequences of actions should be valued. This can be done only by some kind of philosophical analysis into the question of what constitutes "rational conduct."

CHAPTER ELEVEN

TEACHING ANALYTIC SKILLS IN THE SOCIAL STUDIES: A REVIEW OF RESEARCH

The basic goals of political education which constitute the focus of our work are, of course, commonly recognized as critically important for the maintenance of a democratic society. The educational literature is replete with grave pronouncements about the importance of these goals.¹ Obviously this does not mean that the rest of the student's educational development or training in the social studies should be neglected or ignored, but rather that we should perhaps show a rather deep concern for those analytic skills which are, in the long run, essential for the preservation of a society committed to rational consent as the intellectual means by which human dignity can be redefined in each new generation.

Assuming that these analytic skills or abilities are of such central importance to the society, one might expect to find a great deal of curriculum development and research in this particular area of the social studies. Unfortunately not only "new ideas and practices,"² but also research reports on this phase of citizenship training are not plentiful.³ To give some perspective to our own work, however, it is instructive to see what has been done.

Glaser's Study

A classic study in critical thinking is that of Glaser.⁴ Because of its position as probably the first significant piece of educational research in

¹See for example: H.C. Anderson, F.C. Marcham, and S.B. Dunn, "An Experiment in Teaching Certain Skills of Critical Thinking," Journal of Educational Research XXXVIII, 1944, p. 241; Prudence Bostwick and others, "The Nature of Critical Thinking and its Use in Problem Solving," Skills in Social Studies, the 24th Yearbook of the National Council for the Social Studies, edited by Helen McCracken Carpenter, (Washington, D.C.: National Council for the Social Studies, 1953), p. 48; L. Dressel and B. Mayhew, General Education: Explorations in Evaluation (Washington, D.C.: American Council on Education, 1954), p. 35; E.F. Pfeiffer, "Needed Research in Education for Citizenship," Phi Delta Kappan, XXXIII, 1951; S.A. Rippa, "Toward a Definition of Citizenship Education," Social Education, XXIII, 1959; Eugene R. Smith and Ralph W. Tyler, Appraising and Recording Student Progress, Vol. III of Adventure in American Education (New York: Harper and Brothers, 1942), p. 35; G.P. Taylor, "Teaching the Art of Decision-Making," Journal of General Education, VIII, 1955, p. 255.

²K.B. Henderson, "Teaching of Critical Thinking," Phi Delta Kappan, XXXIX, 1958, p. 282.

³A.F. Joyal, "Research in Citizenship Education," Phi Delta Kappan, XXIX, 1947, p. 135.

⁴Edward M. Glaser, An Experiment in the Development of Critical Thinking (New York: Teachers College, Columbia University, 1941).

the area of critical thinking, Glaser's work should be mentioned here although the research was not specifically in social studies.

Glaser used eight twelfth-grade English classes in his experiment. Four experimental classes were matched with four control classes on the basis of average school grades for the previous year, scores on the Otis Quick-Scoring Test of Mental Ability, and scores on the Nelson-Denny Reading Test. The schools contributing the classes grouped students for English instruction into three levels depending on their grade (mark) in English for the previous marking period. Glaser's groups were from the middle level.

The experimental groups studied materials prepared by Glaser specifically to teach critical thinking. The control groups studied the usual course work, but one of the stated objectives for the schools from which the classes were drawn was

the development of ability to think critically. The teachers of both the experimental and control classes, then, were trying to develop in students the ability to think critically. The control teachers, using their own methods, taught the content outlined in the New York and Newark syllabi respectively. The experimental teachers used both their own methods and the procedures suggested in the lesson units. They supplemented the content outlined in the units by other materials which they selected as pertinent to the interests and needs of their students.⁵

Each class was taught by a different teacher. The materials prepared for the experimental teachers included eight "units" including a statement of objectives, basic concepts, materials, suggested procedures, and readings for the teacher. The instruction in these units covered a ten-week period.

The classes were given the Watson-Glaser Tests of Critical Thinking before and after the instruction period. Each of the experimental classes showed a greater average gain on the composite score for the battery of tests than did any control class. Moreover, the average gain on that battery of tests was significantly greater (C.R. of 6.09) for the experimental classes as a group than for the control classes as a group. Glaser concluded:

This significant difference supports the conclusion that the lesson units which have been developed can be used effectively with high school (and probably college) students to stimulate growth in ability to think critically.⁶

⁵Ibid., p. 85.

⁶Ibid., p. 175.

Tests of significance were limited to the differences in mean gain. "The gains of the control groups were not tested to see if they approached significance without special instruction. However, Glaser's investigation did show that greater gains in critical thinking can be obtained by teaching specifically for that end using specially prepared materials. This generalization should probably be qualified in view of the fact that his sample was drawn from the middle grouping of students.

Probably the greatest deficiency in Glaser's study is the lack of theoretical grounding either with respect to the structure of "critical thinking" or in the area of psychological explanation. Teaching materials and procedures are selected on an ad hoc basis, and predictions about learning outcomes are simply logical hunches. This limits the scientific usefulness of the study because it is difficult to draw any implications beyond the conclusion that teaching specifically for "critical thinking" seems to result in gains in critical thinking. There is no explicit conceptual framework to help account for this success.

College Studies:

The eight year study. Some investigations are included in this review of secondary social studies research only by implication. Among these are three studies primarily at the college level. One of the most famous of educational investigations is the Eight Year Study of the Progressive Education Association which attempted a comparison of "traditional" schools with 30 experimental "progressive" schools. Arrangements were made in 1932 for graduates of the experimental schools for several years to be admitted to most universities and colleges in our country upon recommendation of their high school principals and submission of complete high school records. The schools were "set free by the colleges from the usual program of college prerequisites to engage in experimental study of secondary education."⁷

A staff was set up to evaluate the effectiveness of the new educational programs. One of the specific objectives was the teaching of critical thinking.⁸ However, the evaluation was directed at pupil progress, not at a comparison of outcomes in the traditional and progressive schools.⁹

The central question of the study was whether or not the students from the progressive schools would do relatively better than other students in college. A follow-up in several universities and colleges was undertaken. The results were inconclusive although slight differences were found in favor

⁷Dean Chamberlin and others, Did They Succeed in College?, Vol. IV of Adventure in American Education (New York: Harper and Brothers, 1942), p. 1.

⁸Eugene R. Smith and Ralph W. Tyler, Appraising and Recording Student Progress, ch. 2.

⁹Ibid., p. 154.

of the experimental students in terms of absolute achievement in relation to scholastic performance, performance on placement tests, probationary actions, academic honors, and graduation.¹⁰ However, Lancelot¹¹ suggests that it may not be appropriate to attribute even the slight differences to "progressive education." He points out that the ability of the teachers and pupil-teacher ratio in the respective schools might be a better predictor. In doing so, he points out the primary weakness of the investigation -- the difficulty of experimental control with 30 participating schools.

The cooperative study of evaluation in general education. This study was carried out under the auspices of the American Council on Education. One of the committees set up as a part of the project was the Intercollege Committee on Social Science Objectives. This committee selected critical thinking as the basic skill of citizenship and developed a Test of Critical Thinking in Social Science as an evaluation device. (See Chapter Ten.)

With several colleges participating, the test was used in a number of situations to "study growth and achievement in social science, as measured by the Test of Critical Thinking in Social Science...."¹² The various studies are not thoroughly reported, nor were they intended to be. The investigations were meant to be of an exploratory nature "raising some highly provocative questions on which future research is needed" and focusing "attention on the significance of the kind and quality of collegiate teaching."¹³ The results are summarized in the following manner:

This test was administered to a great many entering freshmen at a number of participating colleges, and to these same students or to comparable groups at the end of the freshman year and at the end of the sophomore year. The changes in test scores over these time intervals were studied with a view to determining their magnitude and the factors associated with them. In general, it was found that students gained in ability to think critically in social science over a period of a year, although the size of these gains varied widely, depending on the institution that students attended. Attempts to teach critical thinking in social science by making minor changes in particular courses did not appear to result in greater growth than was found in courses not making overt attempts to teach this skill. Attempts to relate growth in critical thinking ability to course organization or to

¹⁰Chamberlin and others, Did They Succeed in College?, ch. 1.

¹¹W. H. Lancelot, "A Close-up of the Eight-Year Study," School and Society, LVIII, 1943.

¹²L. Dressel and B. Mayhew, General Education: Explorations in Evaluation, p. 56.

¹³Ibid., pp. 64-65.

specific teachers suggested that both of these were highly important, although the research could not identify specific factors which seemed to be operative.¹⁴

Dressel and Mayhew also point out:

The evidence shows that there is growth, over a period of a year, and over a period of two years, with respect to the traits assessed by the test of Critical Thinking in Social Science and that different groups grow at different rates. This suggests that it is possible to plan learning situations so as to facilitate development of critical thinking ability. How this might be accomplished is still largely conjectural, although some clues have been provided.¹⁵

In general terms, Glaser's finding is supported; it does seem that the way to teach critical thinking is to plan and teach specifically for that outcome. Beyond that, the implication from the study is that a unifying conceptual framework is needed to relate "factors." One reason for inconclusive results from the study was probably the lack of a theoretical framework to guide research efforts and the evaluation of these efforts.

Critical thinking in general psychology. Lyle reported a study on the college level "designed to throw some light on the question of classroom methods as related to changes in critical thinking abilities in students."¹⁶ The changes to be evaluated were those in scores on the ACE Critical Thinking test, Form g. (See Chapter Ten.) Two groups of 27 and 28 students were used with the same instructor teaching each group. The groups were not matched, but "approximately equal mean scores were obtained for the two groups on the critical thinking test, administered at the beginning of the semester and on a 40-item multiple-choice pretest of achievement in general psychology." However, "on the American Council on Education Psychological Examination there was a difference of ten points between the percentile ranks of the mean scores for the two groups, the control group having the higher mean score. The two groups were fairly equal in class size, sex ratio, proportions from various college classes, and proportions pursuing various college majors."¹⁷

Two different course procedures were used for the experimental and the control groups. The experimental group picked out problems in psychology to

¹⁴Ibid., pp. 66-67.

¹⁵Ibid., p. 65.

¹⁶Edwin Lyle, "An Exploration in the Teaching of Critical Thinking in General Psychology," Journal of Educational Research, LII, 1958, p. 129.

¹⁷Ibid., p. 130.

be studied. The course was organized around these problems, with discussions, committee and individual reports, and a term paper which was to be the report of an investigation of a problem. In the control group, the teacher set up the objectives. The classes were largely "lecture-discussion." However, because of scheduling,

it was possible from time to time to include in the lectures for the control section a statement of some problems which had been discussed in the experimental sections. However, the method was different between the two sections in that students in the experimental section were given an opportunity to discuss the implications of the problems and arrive at their own conclusions; whereas, the instructor simply recited certain possible conclusions to students in the control section.¹⁸

The improvement in critical thinking scores for the two groups was about the same. The mean gain was 3.4 points for the experimental section and 3.8 points for the control section. Lyle reported that the data suggested that students of higher academic ability did better in the experimental group in terms of improvement, while low academic students did better in the control group. But he reports no statistical tests of this relationship.

An Elementary School Study

Hiram¹⁹ reports a study in the teaching of critical thinking with elementary school students as subjects. The study is included in this review since Hiram's subjects were "upper grade" children. Hiram developed his own test of critical thinking. He then tested 200 students on mental age, general reading ability, general language proficiency, and initial reasoning ability. From the total number of students, he selected two groups of 33 each, equated on the measured variables. He taught the experimental group himself, while nine teachers taught the portions of the control group which were selected from their respective rooms. The experimental period of teaching was four months of 250 minutes per week. The control students were taught the regular course work, which included neither the content nor method of the experimental group.

The objective of the investigation was to test the hypothesis that critical, i.e., logical, thinking can be taught best by teaching children the rules of logic by "direct instruction" consisting of: "a. Materials and learning content which embody the principles of logic. b. Teaching methods that provide full opportunity for the pupil to discover for himself these principles and to formulate them as generalizations."²⁰ He goes on to point

¹⁸Ibid.

¹⁹G. A. Hiram, "Experiment in Developing Critical Thinking in Children," Journal of Experimental Education, XXVI, 1957.

²⁰Ibid., pp. 125-126.

out that these were developed for the purposes of his research with

the former acquired through an analysis of the understandings deemed necessary to an adequate concept of the nature of thinking; the development of the latter grew out of the theory that the most meaningful learnings as well as the most transferable ones are those acquired through self-discovery on the part of the learner and formulated by him into generalizations. Thus the method employed in the experimental instruction was essentially the Socratic Method and consisted of carefully phrased questions based on simple problems. These questions were designed to lead the pupil to discover a pattern in a series of such problems. Once the pattern had been discerned, the pupil was encouraged to express his insight in the form of a generalization. The method of teaching, then, was primarily inductive in nature.²¹

The students in the control and experimental groups were tested before and after the instruction period with Hiram's Test of General Reasoning Ability. The difference in the mean achievement of the groups was tested for statistical significance. The result was a Critical Ratio of 14.57 in favor of the experimental group. The final scores of the experimental group on the reasoning test were correlated with intelligence scores, mental age, language level, and reading level; and the coefficients were not significant. So, it was concluded that these variables were not factors in the increase in critical thinking as measured.

From all appearances, Hiram's experiment was carefully executed and reported. Moreover, theoretical considerations were made and specific hypotheses drawn from them. His theoretical structure could have been expanded, however, allowing for both a more detailed theoretical description of the "Socratic Method" and of the relationship between the method and predicted learning. However, when one takes into account the limitations of reporting research in journal articles, his attempt was certainly adequate.

Probably the greatest defect in the experimental design of the research under discussion was its failure to expand the existing knowledge about teaching critical thinking. As the research stands, it adds weight to the conclusion drawn from Glaser's study, that if critical thinking is a desired outcome, one should teach specifically for that end. In a sense, it was a replication study as both were concerned with teaching rules of logical thinking. However, different methods of teaching critical thinking could have been used to probe the question of what methods are most effective for this purpose. We have added evidence from the experiment that "critical thinking" can be taught, but we still do not know precisely the range of thinking affected, to what extent the results were due to the particular method used, or how much the results are affected by simple exposure to materials of instruction.

²¹Ibid., p. 127.

Myram recognizes himself the need for experimentation varying the method of instruction, while holding the content constant.²²

A Multi-Subject Study

Henderson²³ reports an experiment involving 36 teachers and about 1500 students in English, geometry, science, and social studies. Although the results are not broken down by subject area, the report is reviewed here because social studies classes were included in the sample. The investigation was

designed to test the hypothesis that if knowledge related to the problems of (1) determining the meaning of an expression, (2) determining whether or not an argument is valid, (3) deciding whether a statement is true or false, and (4) justifying opinions and evaluating other people's justifications of their opinions is taught, students will improve in ability to think critically.²⁴

One year was spent in developing and testing instructional materials and one year in the actual instruction with the materials. "Important variables" were held constant as far as possible by matching and by techniques of covariance. The Watson-Glaser Critical Thinking Appraisal Test and the ACE Test of Critical Thinking, Form G, were the evaluation instruments. The difference in mean gain between the experimental and the control groups was 2.15 on the Watson-Glaser Test, and was significant at the .05 level. The difference in mean gain on the ACE Test was not significant. Samples of students from the experimental and control groups took "a free response test involving evaluation of the argument in a fictitious letter to an editor." No mention is made of scoring techniques or controls, but a difference of 2.21 -- significant at the .05 level -- favoring the experimental groups is reported. There was no significant difference in gains on tests of conventional subject matter.

Henderson specifically mentions that he is making a brief report of an experiment to be published in full. The report reviewed is of little value, except that one is probably safe in drawing the conclusion that the Glaser study is once again confirmed. Specific teaching for "critical thinking" seems to produce results. However, the results are inconclusive with conflicting findings on two measures. In general, the report is inadequate for any but the most general of conclusions, and it is only for that purpose that it is included here.

²²Ibid., p. 130.

²³K. B. Henderson, "Teaching of Critical Thinking."

²⁴Ibid., p. 281.

Secondary School Social Studies Research

An experiment with "slow learners." Miller and Weston²⁵ report an attempt to teach a group of tenth grade geography students of low intelligence to think critically. It is indicative of the tone of the entire report that no mention is made of either the number of students involved in the experiment or their scores on a test of intelligence.

The pupils picked out problems, e.g., "Why should the governments of the United States and Canada demand that communities prevent water pollution?" They then proceeded to gather data, and arrived at conclusions, using discussions as a method of presenting data and drawing conclusions.

Evaluation of learning outcomes is primarily on the basis of the teacher's observations. As far as can be ascertained from the report, no attempt was made to control or systematize these observations, as is indicated by reports such as, what "we saw...." indicated "growth." However, J. W. Wrightstone's Test of Critical Thinking in Social Studies was also used to measure growth. It is mentioned that the experimental group improved more than an undefined control group on only one part of the test -- Part II dealing "with ability to draw conclusions." No statistical test of the difference in gain is mentioned.

The results of the investigation under discussion are inconclusive mainly because the gains in critical thinking were not clear-cut. The experimental group improved an undefined amount more than the control group on one part of the Test of Critical Thinking in Social Studies, and this lack of more explicit gain is apparently inconsistent with the teacher's intuitive judgments. Probably the greatest value of the report is that it points up the need for careful research in this area of the social studies curriculum.

The Stanford Social Education Investigation. Three "approaches" to the teaching of social studies that are often mentioned in the consideration of curriculum problems are the "chronological," the "topical," and the "problems" approaches.

The participants of the Stanford Social Education Investigation believed, therefore, that it was important to set up a research study in which the merits of the three approaches could be tested so that inferences might be drawn from the data which would have value for curriculum organization and classroom procedures.²⁶

²⁵J. Miller and G. L. Watson, "Slow Learners Improve in Critical Thinking," Social Education, XXII, 1958.

²⁶J. I. Quillen and L. A. Hanna, Education for Social Competence (New York: Scott, Foresmann, 1948), p. 141.

It was agreed "that the superiority of one approach over the others should grow in the behaviors considered desirable by social-studies teachers, regardless of the approach used."²⁷ Among the "behaviors" defined as desirable objectives of social-studies teaching were the following:

- A. Development of the ability to think critically.
 A person who has this ability:
 1. Interprets data accurately
 2. Uses logical arguments
 3. Recognizes fallacies and sophistries in the arguments of others
 4. Draws sound conclusions
 5. Applies accepted principles to new situations.²⁸

The sample for the experimental teaching involved only eleventh and twelfth grade students because:

- (1) better techniques and instruments of evaluation have been developed on the senior than on the junior high school level;
- (2) more of the teachers participating in the Investigation at those grade levels were enthusiastic about the problems approach and were skilled in using it; and (3) the comprehensiveness of the evaluation program contemplated made it necessary to confine the study to one or, at the most, two grade levels.²⁹

The sample was drawn from five schools -- all "relatively" large public schools, with student bodies similar in race and nationality (native born white), and with a high percentage (36 to 70 percent) of graduates that go on to college. There was some variation in pupil-teacher ratio (24.6 to 34.2) and in the cost of education per pupil (\$95.00 to \$197.00), both of which might well be important variables in predicting the attainment of educational goals such as those involved in this study. However, the researchers concluded:

Since, however, in one case two or three teachers from the same school participated in this study, one using the problems approach and another using the topical or chronological approach, it is unlikely that this variation affected the results of the study in any measurable degree.³⁰

²⁷Ibid.

²⁸Ibid.

²⁹Ibid., p. 143.

³⁰Ibid.

It is doubtful that this conclusion is warranted. It is particularly difficult to estimate the effect which pupil-teacher ratio and cost per pupil might have had because no data is provided as to the actual distribution of the experimental "approaches" among the various schools with respect to these two variables. This leaves unresolved an ambiguity that must be kept in mind in interpreting the results of the investigation.

The teachers participating in the study were volunteers, selected from a larger group so that they were comparable "in teaching load, academic and professional training, teaching experience, and teaching skill."³¹ However, the teachers using the topical approach actually averaged 6.9 years more experience than those using the problems approach with whom they were paired. Nine teachers were involved in the study. Four taught on the eleventh grade level, and five taught twelfth grade students. Four teachers used the problems approach, three the chronological approach, and two the topical approach.

After eliminations because of the usual experimental hazards, such as missed tests, the sample included 465 students. Of these, 270 were seniors and 195 were juniors. The senior groups using the topical and problems approaches were highly comparable in intelligence (Otis Self-Administering Test, Higher Form), reading ability (Iowa Silent Reading Test), and socio-economic status (Wrightstone's Social Background Data Sheet). However, they were not so well matched on age and sex. The problems group was on the average about six months older, and the problems group was about three-fourths boys while the topical group was about one-half boys. The junior groups using the chronological and problems approaches were well matched on intelligence, but not so well on reading ability (problems group tended to score higher), socio-economic status (chronological group tended to be higher), and chronological age (problems group was older). With both seniors and juniors the only difference that was significant beyond the .05 level was chronological age.

One school year was the period of treatment. The assessment instruments were the Interpretation of Data Test, the Social Problems Test, and the Nature of Proof Test, all developed by the Evaluation Staff of the Eight Year Study. The method of evaluation was to compare the number of areas in which the paired groups made significant gains, with a Critical Ratio of three the cut-off point. In comparing the groups taught by the problems and the topical approaches, the researchers concluded:

The students in the groups using the problem-solving approach not only made significant growth in more aspects of critical thinking than the students in the groups using the topical approach, but they also demonstrated superior ability.³²

³¹Ibid., pp. 145-147.

³²Ibid., p. 160.

In other words, the problems group made more significant gains in growth in the ability to think critically as measured, and also had higher mean scores on the final tests.

In comparing the groups using the problems and chronological approaches, the conclusion was:

The students in neither of the junior groups using the problems or the chronological approach made growth which was significant in many of the aspects of critical thinking evaluated in the study.³³

No pattern of difference in growth emerged with each group making some growth, but neither showing significantly increased mean scores on many of the sub-tests.

The results of the study are thus quite inconclusive. The experimental design provided for no comparison of the chronological and the topical approaches. While the data seems to suggest that the problems approach is more effective than the topical approach in teaching critical thinking, no such suggestion of differential effectiveness can be drawn from the data comparing the problems and chronological approaches.

It should be pointed out that the study used an extremely primitive theoretical grounding. There was only a meagre attempt to describe the teaching approaches in terms of content organization and orientation. No account was taken of teacher behavior in using the approaches, and no attempt was made to formulate a theory relating the approaches to possible learning outcomes.

The study thus provides us with little general knowledge of the effectiveness of methods of teaching critical thinking in the social studies. Once again, however, it should be noted that the approaches were not specifically aimed at the teaching of critical thinking, although it was an objective. It was hoped that critical thinking would be a by-product of method.

"Doing" and "telling" methods and critical thinking. An experiment in teaching specific critical thinking skills to social studies students is reported by H. C. Anderson, Marcham, and Dunn³⁴ as a part of the work in the teaching of critical thinking carried on at Cornell University. The sample of students included 14 pairs of classes in eight Iowa and New York cities at the seventh grade level, and 12 pairs of classes in eleven cities in the same states at the tenth grade level. There were, for the purposes of final comparison, 715 students in the seventh grade group and 597 students in the eleventh grade group.

³³Ibid., p. 171.

³⁴H. C. Anderson, F. C. Marcham, and J. B. Dunn, "An Experiment in Teaching Certain Skills of Critical Thinking."

The paired classes were assigned to be taught by the "doing" or the "telling" method by random numbers in order to avoid selection bias. Five "problems" were developed for use at each of the two grade levels, with the problems at each grade level varied as to content and reading difficulty. Each of the problems contained enough material for about three class periods of instruction. Each problem was in two forms, "doing" and "telling."

In the latter case, the experimental material was essentially self-sufficient. That is to say, a reading section was provided, and the way in which an expert would proceed in analyzing this content was explained in detail. Problems organized according to the "doing" method contained the same basic selection, but pupils were left comparatively free to study and analyze this content as they saw fit. In the teacher's manual, however, recommended procedures were suggested for the "doing" type of problems.³⁵

Two "objective-type" tests were used to measure the gains in the skills stressed by the experimental materials: Test I, Abstracting and Organizing Information and Test II, Drawing Conclusions. The result was that with the groups pooled on the basis of teaching method, there were no significant differences in gains in critical thinking ability as measured by the two tests. However, within both methods there was a great variation in results, "with some pairs of classes favoring the 'doing' method and some the 'telling' method, at both grade levels."³⁶ The variation is explained as follows:

The observations of pupil and teacher reactions made by the research assistants during the course of the year suggest a general explanation for this inconsistent reaction. The more imaginative teachers and the abler pupils showed greater interest in the 'doing' method; the less imaginative teacher, or the one who devoted little time to the preparation of lesson plans, found it much easier to use the 'telling' method. In many cases, slow pupils seemed to derive greater benefits from the more explicit 'telling' materials.³⁷

This last statement reveals what is probably the greatest contributor to the inconclusive results of the study -- the failure to control either experimentally or statistically for student intelligence or to describe carefully teacher behavior. The report does mention the use of analysis of covariance. However, the only results shown in the tables are the mean gains and the differences between mean gains for the paired groups. No mention is made of levels of significance or adjusted means which might imply that the technique of covariance had actually been applied. No levels of significance were reported.

³⁵Ibid., p. 242.

³⁶Ibid., p. 249.

³⁷Ibid.

Elements of the Experimental Situation

As indicated in Chapter Eight during the first year of the Project a great deal of experimentation was done with the development of teaching materials, teaching styles, evaluation instruments, and personality measures. At the end of the first year it was decided that we had enough confidence in our hunches about the development and evaluation of an experimental curriculum that we should give such a curriculum a fairly careful trial. There were numerous times after this decision was made that it seemed premature. In retrospect, however, the pressure provided by the necessity of setting up such an experiment produced a rapid clarification of many ideas which had not previously been rigorously analyzed or specifically stated. The point is simply that this was a first systematic effort: many modifications and much cutting and fitting took place during the experiment itself which violated the concept of "pure research." Moreover the design which was set up to compare the experimental students with control students had weaknesses, mainly because of the resources available and the scheduling problems inherent in any public school situation. Despite these difficulties, however, we feel the results are interesting and useful to anyone interested in our general approach to the problem of teaching secondary school students controversial political issues in the schools.

The Students

The experimental curriculum was initiated when the students entered the junior high school in grade seven. Boys and girls are approximately 12 years old at this point in their academic career. The general characteristics of the community described in Chapter Eight give some idea of the type of student with whom we were dealing. His average I.Q. was approximately 113; with some exceptions he was interested in doing well in school; and while the backgrounds of the students varied, the great majority came from middle class homes.

The Curriculum

While the materials developed for the experimental curriculum were taught over a two year period, the teaching time required for the EC materials themselves was only a little over a year. These materials were "fitted" into the chronological sequence of a standard American history curriculum. Allowing for the interruptions common to any public school routine, approximately one-third of the student's time was occupied with textbook American history, slightly more than one-half of the student's time was occupied with the EC, and the remainder consisted of place geography, current events, vocational guidance, taking standardized tests for school records, and a great deal of testing connected with the Project. Table 11.1 will give the reader some indication of the overall time-table of the two year period. The table does not account for some 13 weeks which were taken up by the miscellaneous activities indicated above as well as a wide variety of other shorter activities too numerous to mention.

TABLE 11.1

TIMETABLE DESCRIBING SOCIAL STUDIES PROGRAM FOR
STUDENTS ENGAGED IN TWO YEAR EXPERIMENTAL CURRICULUM (GRADE SEVEN)

Topic	Project Activities	Non-Project Activities
Europe Discovers and Explores the New World to the West		3*
European Nations Develop Colonies in the New World		3
Pretest Achievement Battery	2**	
Critical Thinking Skills with Short Illustrative Cases	8	
Birth of the American Republic	1	1
An Introduction to the Structure and Principles of American Government	5	
Application of Model for Handling Political Controversy Using Analytic and Political Concepts	3	
Establishment of the New Nation		3
Economic Changes in the New Nation and the Rise of Sectionalism		3
First Post-Test Achievement Battery	1**	
Total Weeks-Grade Seven	17	13

*Numbers indicate weeks

**Not included in total

TABLE 11.1 -- (Continued)

TIMETABLE DESCRIBING SOCIAL STUDIES PROGRAM FOR
STUDENTS ENGAGED IN TWO YEAR EXPERIMENTAL CURRICULUM (GRADE EIGHT)

Topic	Project Activities	Non-Project Activities
Second Post-Test Achievement Battery	2**	
American Civil War and Reconstruction		3
School Desegregation	6	
Settling the West		1
The American Indian	1	
Introduction to the Industrialization of America		1
The Problem of Fair Competition and Business Monopoly	4	
Vocational Guidance		1
American Immigration Policy	1	
Problems of American Labor	4	
Personality Test Battery	2**	
United States History: 1900-1925		2
The New Deal	3	
United States History: 1940-1960		2
Review	1	
Third Post-Test Achievement Battery	2**	
Total Weeks-Grade Eight	20	10
Total Weeks	37	23

**Not included in total

Table 11.2 should give the reader a more specific idea of the information, concepts, and documents as well as their sequencing which constituted the EC, and the way it was integrated into the standard American history course. It is obviously not feasible to present all the materials. Each document will, however, be described in brief. This, along with the unit outlines and appendices to Chapter Nine should give the reader a clear idea of the type of content the student encountered.

We might note here the overall organization of the EC as it was initially set out. It was organized in three distinct parts:

- (1) analytic material (described in Chapters Five through Seven);
- (2) background, principles, and structure of American government (partly summarized in Chapters Three and Four); and
- (3) a series of substantive problem units to which the analytic concepts contained in (1) and (2) were to be applied.

After teaching parts (1) and (2), however, we found that the two sets of concepts were organically related, and that the analysis of values described in the critical thinking material required the use of governmental values and principles, and an understanding of how these principles might be construed. (See Chapter Two.) It was at this point that we decided to provide the students with experience in handling the model integrating the governmental principles with the critical thinking concepts. This occurred in the latter part of grade seven. The second year of the program was implemented very much as it had been planned: the students were taught to apply the analytic model to a series of controversial cases embedded in broader issues, organized generally on a topical basis. It is also important to note that all documents in the curriculum were supplemented by practice exercises and brief illustrative cases developed on a daily ad hoc basis.

Teachers and Teaching Style of the Experimental Program

In setting up an experimental curriculum, there are two major considerations: the materials one intends to present to the students, and the approach the teacher intends to take in handling the material in the dialogue with the students. In this latter area, it is impossible to talk about a single approach, since there were several categories of teaching material. Historical and conceptual background material was handled in a straightforward descriptive didactic way. The material was amplified and explained by the teacher, student questions were answered, and the students were asked questions to test their understanding of the material. Emotional impact or empathy materials required very little "teaching." The materials carried the weight of their own message. After the reading of such a document the teacher would simply try to phrase the problem raised by the case in more general terms: How does the problem which affects this individual or group also affect the whole community? Take off materials, i.e., materials used mainly for independent study by the students, require individual guidance by the

TABLE 11.2

A SUMMARY OF THE EXPERIMENTAL CURRICULUM

UNIT ONE: CRITICAL THINKING		(First exposure, approximately 8 weeks)
<u>Document</u>	<u>Type</u>	<u>Description of Content</u>
"Learning to Think Critically"	A Text	I. Describing the World Around Us
		A. Definitions: pointing and criterial definitions
		B. Classes and classification 1. Danger of classifying an object on the basis of limited information 2. The danger of using classes which refer to averages
		C. Definitions and terms with value loadings
		II. Testable Statements: Claims
		A. Statements that describe events in the world around us 1. Specific claims 2. Summarizing statements - Generalizations - Statistics 3. Explanations
		B. Telling How Sure We Are 1. Statements which are true beyond a reasonable doubt 2. Statements which are probably true 3. Statements which are false beyond a reasonable doubt 4. Statements which are probably false 5. Statements which are doubtful 6. Statements which are controversial
		III. Proof Process
		A. Framing hypotheses
		B. Assumptions or hidden claims implied by a hypothesis
		C. Sampling: Stating how much evidence supports a claim
		D. Testing complex explanations
		E. Sources of Evidence 1. Intuition 2. Authority 3. Personal observation 4. Proof by analogy

TABLE 11.2 -- (Continued)

UNIT ONE: (continued)

<u>Document</u>	<u>Type</u>	<u>Description of Content</u>
IV. Value Judgments		
A. Value Judgments and Decisions		
B. Statements of preference		
C. Dilemmas		
D. Loaded Statements		
V. Argumentation		
A. Where an argument begins		
B. Two levels of argumentation: belief and attitude		

UNIT TWO: BIRTH OF THE AMERICAN REPUBLIC (two weeks)

<u>Document</u>	<u>Type</u>	<u>Description of Content</u>
Chapters 7-9 in regular textbook <u>This is America's Story</u> Wilder, Ludlum, and Brown (Houghton Mifflin)*	Text	Background of American Revolution and early incidents
Excerpts from <u>Oliver Wiswell</u> K. Roberts (Doubleday) [chapters 4 and 5]	Emotional Impact Case	Dramatic account of the persecution of a loyalist family which did not believe in breaking English law
"Boston Tea Party"	Controversial Case	Dramatic account of the Boston Tea Party
"George Watkins"	Controversial Case	Fictionalized dramatic account of a speech by Sam Adams and a meeting of American radicals. Protagonist is portrayed as a "man on the fence."

*Will henceforth be referred to as text: WLB.

TABLE 11.2 -- (Continued)

UNIT THREE: AN INTRODUCTION TO THE STRUCTURE AND
PRINCIPLES OF AMERICAN GOVERNMENT (five weeks)

<u>Document</u>	<u>Type</u>	<u>Description of Content</u>
A. Basic Concepts of Government		
Excerpts from "Strength of the Strong" by Jack London	Conceptual Back-ground	Fictional description of development of government among stone age man. Illustrates <u>power</u> of group versus individuals; danger of being <u>exploited</u> by leaders of group, once a group gets an organized government.
Excerpt from "Two Years Before the Mast," by Richard H. Dana, Jr.	Conceptual Back-ground	Illustrates excesses of <u>autocratic government</u> which are not checked by other agencies of government.
"The Storm"	Conceptual Back-ground	Illustrates disaster which overcomes a wagon train when members refuse to accept expert <u>leadership</u> of qualified men, and substitute government by "common man."
Synopsis of <u>Animal Farm</u> by George Orwell	Conceptual Back-ground	Illustrates governmental cycle: <u>Autocracy--revolution -- democracy -- autocracy</u> . Also importance of <u>propaganda</u> and distortion of language in maintaining political control.
"Sailwell"	Conceptual Back-ground	A Robinson Crusoe story, showing how a group of shipwrecked people develop <u>governmental procedures</u> similar to American system as a response to specific problems facing an island community.
B. Principles of American Government: Their Constitutional Basis and Implementation		
<u>Federal Textbook on Citizenship</u> by John G. Hervey, U.S. Department of Justice, Washington, 1955	Conceptual Back-ground	Basic Values of American Society: <u>Personal freedom</u> <u>Equality and justice</u> <u>Peace and order</u> <u>General welfare</u> <u>Brotherhood</u>

TABLE 11.2 -- (Continued)

UNIT THREE: (continued)

<u>Document</u>	<u>Type</u>	<u>Description of Content</u>
U.S. History text-book -- <u>The Story of American Democracy</u> Casner and Gabriel, (New York: Harcourt Brace), chapter 7.		Governmental Principles <u>Rule of law</u> <u>Rule by consent of governed</u> <u>Equal protection under law</u> <u>Due process of law</u> <u>Separation of powers</u> <u>Checks and balances</u>
<u>For a written version of essentially what was taught, see Part Two</u>		Constitutional basis of values and principles Mechanics of the consent system: selected provisions of the Constitution.

UNIT FOUR: APPLICATION OF ANALYTIC AND POLITICAL CONCEPTS USING SPECIFIC CONTROVERSIAL CASES

<u>Document</u>	<u>Type</u>	<u>Description of Content</u>
		<u>For most of these cases are based on Supreme Court Decisions</u>
"Lunch Counter"	Controversial Case	A fictitious case of a Negro sit-in, demonstration
"Sidewalk Speech"	Controversial Case	An incident involving free speech, based on a Constitutional Case: <u>Feiner v. New York</u> , 340 U. S. 315 (1951).
"Davis Lumber"	Controversial Case	An incident in which a lumber dealer challenges the constitutionality of Fair Labor Standards Act, based on a Constitutional Case: <u>United States v. Darby</u> , 312 U. S. 100 (1941).
"The House"	Controversial Case	Story about housing discrimination against an American citizen of Chinese ancestry.

TABLE 11.2 -- (Continued)

UNIT FOUR: (continued)

<u>Document</u>	<u>Type</u>	<u>Description of Content</u>
"Flag Salute"	Controversial Case	An incident in which a Jehovah's Witness refuses to allow his child to salute the flag in public school, based on two Constitutional cases: <u>Minersville School District v. Gobitis</u> , 310 U. S. 586 (1940) and <u>West Virginia State Board of Education v. Barnette</u> , 319 U. S. 624 (1943).
"The Job"	Controversial Case	An incident in which California prosecutes a man for helping his unemployed brother-in-law enter the state, based on a Constitutional case: <u>Edwards v. California</u> , 314 U. S. 160 (1941).
"Radios and Buses"	Controversial Case	An incident in which two men contest the right of a bus company to install radios on public buses.
"Smittites"	Controversial Case	Story of a religious sect which obtains special tax privileges because of a broadened definition of "religious property."
"Rabin Case"	Controversial Case	An incident in which a counterfeiter's desk is searched when he is arrested. Does this constitute unreasonable search and seizure? Based on a Constitutional case: <u>United States v. Rabinowitz</u> , 339 U. S. 56 (1950).
"Out of Bounds"	Controversial Case	Story about housing discrimination against a Negro doctor.
"Ota Case"	Controversial Case	A fictional account of a Japanese family relocated from the West Coast at the beginning of World War II, based on a Constitutional case: <u>Korematsu v. United States</u> , 323 U. S. 214 (1944).
"Loudspeaker and the Union"	Controversial Case	An incident in which a city restricts the right of a labor organization to use a loudspeaker to take its case to the public, based on a Constitutional case: <u>Kovacs v. Cooper</u> , 336 U. S. 77 (1949).

TABLE 11.2 -- (Continued)

UNIT FOUR: (continued)

<u>Document</u>	<u>Type</u>	<u>Description of Content</u>
"Hours"	Controversial Case	An incident in which the government restricts a woman's right to work or to be hired for as many hours as she wants, based on a Constitutional case: <u>Adkins v. Children's Hospital</u> , 261 U. S. 525 (1923).
"The Voyage"	Controversial Case	Story of a long-time non-citizen resident of the United States who is refused readmission to the United States after visiting his native country behind the iron curtain, based on a Constitutional case: <u>Shaughnessy v. United States, ex. rel. Mezel</u> , 345 U. S. 206 (1953).

All of the above cases were used both to teach and illustrate concepts and principles of American government as well as to teach the thought process patterns described in Part Two.

PROBLEM UNITS

UNIT FIVE: SCHOOL DESEGREGATION (six weeks)

[See Chapter Nine, Appendix A]

UNIT SIX: THE AMERICAN INDIAN (one week)

<u>Document</u>	<u>Type</u>	<u>Description of Content</u>
Text: W.L.B., Chapter 21	Historical Background	
"Trail of Tears"	Controversial Case	Story of the forced migration of the Creek Indians.
"Hopi Indian"	Historical Background	Brief description of the Hopi Indian.

TABLE 11.2 -- (Continued)

UNIT SIX: (continued)

<u>Document</u>	<u>Type</u>	<u>Description of Content</u>
Excerpt from "The Sun Chief" the Autobiography of a Hopi Indian, Edited by Leo W. Simmons (New Haven: Yale University Press, 1942)	Controversial Case	Story of compulsory education for some Hopi Indians.

UNIT SEVEN: THE PROBLEM OF FAIR COMPETITION AND BUSINESS MONOPOLY (four weeks)

<u>Document</u>	<u>Type</u>	<u>Description of Content</u>
A. Rise of Big Business in America		
"Economic Revolution"; Text: W.L.B., Chapters 22 and 23.	Historical Background	A text to provide the student with some insight into the rapid industrial changes occurring between the Civil War and 1900.
"Freight Rates -- The Long Haul and the Short Haul," excerpt from the <u>Octopus</u> , by Frank Norris.	Controversial Case	A case showing how an arbitrary shipping policy of the railroad works an injury on the farmer.
"Rise in Freight Rates," excerpt from the <u>Octopus</u> , by Frank Norris.	Controversial Case	A case showing how a farmer is ruined by an arbitrary rise in freight rates.
"Andrew Carnegie"	Historical and Conceptual Background	A case study of Carnegie to illustrate how an industrialist might gain great power. Also illustrates a number of important economic concepts, e.g., vertical and horizontal integration of industry.
"Some Notes on Mr. Carnegie's Business Practices"	Controversial Case	Shows some of Carnegie's business practices which were legal but ethically suspect.

TABLE 11.2 -- (Continued)

UNIT SEVEN: (continued)

<u>Document</u>	<u>Type</u>	<u>Description of Content</u>
"South Improvement Company"	Controversial Case	Shows how Rockefeller gained monopoly of the refineries in Cleveland through methods of doubtful legality.
"Standard Oil Again"	Controversial Case	Shows how Rockefeller finally gained control of the oil industry with methods of doubtful legality and morality.
"Standard Oil Aftermath"	Historical Background	A brief text to bring some closure to the Standard Oil material.
B. Some Government Efforts to Legislate Fair Competition in Business		
"A Brief History of Government Efforts to Promote Fair Competition"	Historical Background	A text tracing history of government regulation of business from I.C.C. through the Sherman Act, Clayton Act, and Federal Trade Commission Act.
"The Federal Trade Commission and the Supreme Court"	Controversial Case	Four practices branded as "unfair competition" by the Federal Trade Commission which were later appealed to the Supreme Court. Student is asked whether or not he agrees with the FTC rulings.
"Supreme Court Decisions on the Four Federal Trade Commission Cases"	Controversial Case	Shows how the Supreme Court reversed the FTC in two cases and upheld it in two cases, partly because of its effort to maintain its own jurisdiction over such matters.
C. Case Study of Contemporary Business Practices in the Retail Sale of Gasoline.		
"The Gasoline Station"	Historical Background	General discussion of business organization and control in the oil industry, and the various pressures placed upon the retail gasoline dealer.
"Gasoline Dealers Complain to Congress"	Emotional Impact	Summary of an article published in the <u>New Republic</u> showing arbitrary treatment of retail dealers by large gasoline companies. (June 13, 1955, pp. 8-10).

TABLE 11.2 -- (Continued)

UNIT SEVEN: (continued)

<u>Document</u>	<u>Type</u>	<u>Description of Content</u>
"Gas War"	Controversial Case	Presents problem of cut rate station and efforts of company owned and leased stations to deal with it within the framework of free competition.

UNIT EIGHT: AMERICAN IMMIGRATION POLICY (one week)

<u>Document</u>	<u>Type</u>	<u>Description of Content</u>
"Immigration -- An American Problem"	Historical and Conceptual Background	Description of major waves of immigration into the United States, reactions and discrimination against immigrants, and major legislation affecting immigrants.
Text: W.L.B. Chapter 25		
"Immigration -- 1961"	Emotional Impact	Describes a man of English-Burmese descent, who cannot enter the United States because he comes under the Burmese Quota.
Edward R. Murrow film on immigration	Conceptual Background	Gives opinions of many national leaders on immigration, including Walters, Archbishop Cushing, etc.
"Cuban Refugees"	Controversial Case	Describes problem of Cuban refugees in Miami. Asks, what should U. S. policy be?

UNIT NINE: PROBLEMS OF AMERICAN LABOR -- AN INTRODUCTION (four weeks)

[See Chapter Nine, Appendix B]

UNIT TEN: THE NEW DEAL (three weeks)

<u>Document</u>	<u>Type</u>	<u>Description of Content</u>
"The New Deal"	Historical and Conceptual Background	Describes economic and social conditions of the 1920's leading up to the Crash; conditions after the Depression has begun, and the election of 1932.

TABLE 11.2 -- (Continued)

UNIT TEN: (continued)

<u>Document</u>	<u>Type</u>	<u>Description of Content</u>
"Depression"	Emotional Impact	Describes a series of specific cases of discouraged and needy people during the depression.
"Happiest Man On Earth," by Albert Maltz	Emotional Impact	A short story describing the complete demoralization of an unemployed worker and his exhilaration when he finds employment driving a nitro-glycerine truck.
"Agricultural Adjustment Act"	Controversial Case	This case first describes several incidents where farmers resort to violence to keep their farms. Then it discusses AAA and some problems that result from the government program.
"The Blue Eagle"	Controversial Case	Description of how the NIRA attempted to meet the problems of the depression, and some of the criticisms of the government program.
"WPA"	Controversial Case	Description of the problem that led Roosevelt to consider packing the Court, as well as other alternatives available.
"Court Packing and the Public"	Controversial Case	A dramatic presentation on tape of many of the actual statements made by political leaders as well as by the man on the street.

teacher. In these respects all of the experimental students are treated in as similar a way as possible. Two different teaching approaches were used on the experimental groups in Concord in presenting the controversial case, the most important type of document from the point of view of analytic concept application. The two approaches will be described in detail in Chapter Twelve. We will mention them at this point, however, since they relate to the type of teaching to which the Concord groups were exposed.

The two types of dialogue might be called the descriptive-analytic and the Socratic-analytic. As indicated by the labels, both required the student to use the analytic framework described in Part Two to conceptualize the case under discussion. In the descriptive style the teacher was asked to assure himself that the students had the facts of the case clearly in mind, to spend as much time as needed asking the students to clarify an accurate description of the case. He then proceeded to ask students to describe the case in terms of the analytic framework, to suggest how different positions could be justified within that framework, and to suggest critical questions that might be asked to clarify which of the various positions were more reasonable. The Socratic discussion, on the other hand, focused very little on the factual background of the case, but rather spent more time examining individual student's own personal positions with respect to the issue under discussion. The teacher deliberately challenged the student's position, pointing up, in particular, inconsistencies, and then used the analytic framework to suggest how the student might handle his position more intelligently and what questions he might ask of himself to strengthen his understanding of his and other positions. The Socratic discussion theoretically alternated between a process of argumentation between the teacher and the student (with the student's position in question) and an attempt by the teacher to clarify the general nature of the argument as he was able to describe it with the political and analytic concepts of the framework.

It should be understood, therefore, that the Concord students were all exposed to the experimental materials described above; that all were taught the materials in essentially the same manner (except for systematic differences for controversial cases and uncontrolled differences due to individual teacher styles). During the teaching of controversial cases, regular class groups were split into two matched subgroups. One of these matched subgroups, then, was taught by a teacher using a Socratic-analytic style; the other was taught by a teacher using a descriptive-analytic style. Although teaching style will be discussed in detail in Chapter Twelve it is important to keep this brief description in mind when we compare the experimental curriculum in Concord with conventional curricula in control schools.

The actual implementation of the EC was carried out by four teachers. The educational background and teaching experience is described in Table 11.3. As we pointed out earlier, all four teachers had primary responsibility to the Project, and during the time of the experimental teaching taught only experimental classes.

TABLE 11.3

EDUCATIONAL BACKGROUND AND TEACHING EXPERIENCE
OF THE EXPERIMENTAL TEACHERS

Teacher	Undergraduate Degree	AMT or Ed.M	Teaching Experience Prior to Fall, 1959
B	B.A. in Political Science, Boston University, 1954	AMT Harvard 1957	2 years, junior high school
G	B.S. in Education Boston University, 1950	Ed.M. Harvard 1960	9 years, junior high school
O	B.A. in Psychology, Amherst, 1950	AMT Harvard 1951	2 years, junior high school
S	B.A. in History, Amherst 1955	AMT Harvard 1959	$\frac{1}{2}$ year, junior high school

Design

It should be emphasized again that the experimental design involved an experiment within an experiment. First, learning gains of the Concord students who were taught the two year experimental curriculum were compared with the gains of students in three control schools using more conventional curricula. Second, analyses were carried out using only students in the experimental program. In this second phase, we investigated the effect of two teaching styles upon the ability of students to handle controversial cases, the possible effects of different experimental teachers working from the same basic curriculum, possible relationships between student personality variables and the effectiveness of the curriculum, and possible interactions between student personality variables and Socratic versus descriptive teaching. The design of the first experiment was very simple, and involved comparing the experimental group with groups of students in three other schools. Some of the more relevant characteristics of these schools are described in Tables 11.4 and 11.5. The design for the more intensive research within the experimental curriculum on style differences, teacher differences, and student personality differences will be presented in Chapter Twelve. At this point we should note, however, three important considerations.

1. Students in the experimental program knew they were being treated in a special way. They were subject to special testing. Their discussions were often recorded. They were aware that four teachers, three of whom were directly affiliated with the university, were spending a great deal of time developing this curriculum. This certainly may be looked at as a systematic advantage for the experimental group.
2. The four teachers involved in the Project were selected both because of their interest in the program and its specific objectives. And although their educational experience and training probably do not differ significantly from the average teacher in a suburban school system, their general competence as well as their interest in contemporary affairs and educational innovation undoubtedly does.
3. For approximately half of the experimental program, regular size classes -- about 25 students -- were split in order to provide an adequate number of groups for experimentation in phase two. This means that about half of the teaching was done in student groups of approximately twelve. This, of course, is an important qualification of all results.

It should be pointed out that the experiment was looked upon as a "hothouse" model.* We wanted first to establish optimum conditions for teaching, and then study variations in learning. This is the reason that we selected the highest quality teachers we could find. Of course, the splitting of groups was based primarily on research design needs, but also was acceptable because it optimized group size for instruction. We had little doubt, moreover, that experimental students would show greater gains than control students

*We realize that "hothouse" experimentation is vulnerable to the criticism: How can you interpret your findings if you don't have a design complex enough to take into account systematic (but theoretically unimportant) differences between the experimental and control groups. Our feeling has been that the major objective of this research was to maximize the possibility of differences between control and experimental groups, but maintain experimental rigor between the two experimental groups. We really were not interested in the extent to which experimental gains were due to the enthusiasm or competence of the teachers or the advantages of small group teaching. We assumed that these may be necessary conditions for experimental gains, but that inclusion of the curriculum itself would also be a major necessary condition. We clearly have a biological analogy in mind: people with appendicitis may do slightly better by having enthusiastic and intelligent doctors and nurses tend them, by being in private rooms rather than large wards, etc., but there is a more basic treatment (an appendectomy) which makes the critical difference. Nor would one want to demonstrate the effectiveness of this particular medical operation in a dirty operating room with mediocre surgeons. McKeachie, who has had much experience in classroom experimentation, makes a similar point:

Generally I would not worry too much about the presence of uncontrolled variables which may account for positive results if there are variables in teaching. Thus you need not be too concerned at this preliminary stage over the possibility that faulty sampling or biases in test scoring are responsible for positive results. In your experiments go ahead and use teachers who are enthusiastic about the new methods you want to use. Go ahead and sell your students on the great advantages of being in an experimental group. Your most probable outcome is "No significant difference." If you come out with a significant difference, you at least have shown that your measures are sensitive enough to show differences and that under ideal circumstances you can improve on what was done in your control group. With this as a basis, you are in a much better position to go on to pin down specific variables than if your original results had been negative. The only good effect of negative results is to squelch some of the wildest claims of the enthusiasts for the latest fad.

W. J. McKeachie, "Problems and Perils in Controlled Research in Teaching," Needed Research in the Teaching of English, Proceedings of a conference held at Carnegie Institute of Technology, prepared by Erwin R. Steinberg, 1962, p. 55.

TABLE 11.4

GENERAL CHARACTERISTICS OF EXPERIMENTAL AND CONTROL GROUPS

School	Location	Approx. School Size	Grade Levels in School	I.Q.*	
				Mean	S.D.
Concord (experimental)	Suburban Boston	400	7-8	113.3	12.3
West	Suburban Boston	400	7-8	118.7	14.6
Big	Urban Boston	1,000	7-9	109.8	10.7
Heat	Suburban L.I. (New York)	1,200	7-12	116.7	10.9

*In this study a number of different I.Q. measures were included because tests ordinarily administered by the schools were used. This assumes comparability of I.Q. measures, which may not be too unreasonable if one takes into account the level of reliability commonly obtained from the readministration of a single measure.

TABLE 11.5

BRIEF DESCRIPTION OF CURRICULA IN CONTROL SCHOOLS

School	Curriculum	
	Seventh	Eighth
West	World Geography	United States History
	Latin America 18 weeks	Structure of course is based on <u>This is America's Story</u> by Wilder, Ludlum & Brown (Houghton-Mifflin)
	Asia 12 weeks	
	Africa 6 weeks	
Big	United States History from the discovery of America through the American Revolution	United States History from the forming of the new government through the Spanish-American War
	Major texts used in this two year sequence are <u>This is America's Story</u> and <u>History of a Free People</u> by Bragdon and McCutchen (Macmillan).	
Wheat	Two year sequence in United States History	
	Special features in seventh grade:	Special features in eighth grade:
	Unit on geographic concepts.	Units on state and local history.
	Unit on economic concepts.	
	Anthropological unit on the American Indian.	
	General comments: Emphasis on cultural history relying heavily on biography. No single text used.	

on certain kinds of measures. Glaser's work demonstrated this rather conclusively and several studies have corroborated this finding. The focus of the broad experiment comparing Concord students with control students was directed, therefore, more at the question of what particular kinds of measures might best reflect the special experiences provided the students by the experimental curriculum. The experiment was, in effect, as much a test of the sensitivity of measures of doubtful validity as it was a test of the experimental curriculum.

Aside from the three "uncontrolled" differences between experimental and control groups listed above, our major hypothesis was very simple:

It is possible to teach students in early adolescence to use an abstract conceptual system for analyzing political controversy. Moreover, the ability to use such a system of analysis will (or will not) be reflected in their performance on various measures of critical or analytic thinking.

The Testing Program

The following measures were used in the experiment to explore the basic hypothesis. Although many of these measures were critically reviewed in Chapter Ten, they will be briefly described again so the reader can more conveniently interpret the findings in the next section.

Measures of Problem Analysis: Standardized Tests

1. The Wagnis Test. Three subtests of the Watson-Glaser Critical Thinking Appraisal, Form Am, and two subtests of the Michigan State Test of Problem Solving, Form A, were selected and combined into a single test called the Wagnis Test. The subtests used were: Watson-Glaser Inference Test* which presumably measures the students' ability to determine the level of certainty with which a number of statements can reasonably be inferred from an initial item of information. This test presents statements which may be logically excluded from or implied by an initial message; other statements demand a calculated, though uninformed, inductive inference. The following item is illustrative:

The first newspaper in America, edited by Ben Harris, appeared in Boston September 25, 1690, and was banned the same day by Governor Simon Bradstreet. The editor's long fight to continue his little paper and print what he wished marks an important episode in the continuing struggle to maintain a free press.

*Indicates test used in Wagnis Test.

On the basis of this information, the student is then asked to make the following statements as definitely true, partly true, insufficient data, partly false, or definitely false.

1. The editor of the first American newspaper died within a few days after his paper was banned.
2. Governor Bradstreet felt he had the legal authority to ban Ben Harris' paper.
3. The editor of this paper wrote articles against taxes of the kind which later brought about the "Boston Tea Party."
4. Ben Harris was a man of persistence in holding to some of his interests and convictions.

Watson-Glaser Recognition of Assumptions Test.* In this test the student is given a number of statements each of which is followed by several proposed assumptions. The student is asked to decide whether or not each assumption is necessarily taken for granted in the statement.

The following item is illustrative:

Statement: "Even if all the wealth in the country suddenly were to be distributed equally, some people soon would again become rich and others poor."

Proposed Assumptions

1. The real causes of wealth and poverty would not be much affected by such Socialism.
2. Our present economic system is better than such Socialism.

The student marks these proposed assumptions as "assumption made" or "assumption not made."

Watson-Glaser, Part Three, is a test in syllogisms, and was not used in our evaluation. Considering the limitations on our total testing time, we decided that this particular intellectual process, or at least the way it was evaluated in the test, was probably not useful in analyzing controversial political issues. The following item, perhaps, will indicate the reasons for our reservations about this particular type of test.

Some cannibals are sincere idealists. All cannibals are fanatics. Therefore --

1. Some sincere idealists are fanatics.
2. Some fanatics are sincere idealists.
3. No fanatics are sincere idealists.
4. All fanatics are cannibals.

The student is then asked to discriminate between those statements which do or do not logically follow.

Watson-Glaser, Part Four,* requires the student to state whether or not a number of conclusions drawn from an initial statement follow beyond reasonable doubt.

The following item is illustrative:

A report of the U. S. Census states that during 1940 there were approximately 1,656,000 marriages and 264,000 divorces granted in the United States.

Which of the following statements follows beyond reasonable doubt?

1. Getting a divorce is a quick and easy matter in the United States.
2. If the above ratio still holds true, then about six times as many people get married each year as get divorced.
3. The divorce rate in the United States is much too high.

Watson-Glaser, Part Five, is called Evaluation of Arguments and was not used in the study because of its doubtful face validity. Our conclusion was that arguments were cited as "good" and "bad" because of the implicit personal preferences of the test maker, rather than because of the quality of any intellectual operation involved. The following item illustrates our point:

The question is posed: Should infants be fed by regular schedule rather than whenever they seem hungry?

The Michigan State Test of Problem Solving contains seven parts, and is, in many respects, like the Watson-Glaser Appraisal. Part One is a mixture of items calling for problem identification and assumption identification. Part Two involves syllogistic reasoning; in Part Three the student is asked to identify statements which support a conclusion presented in an initial paragraph; Part Four asks the student to translate statistical information into generalizations in sentence form; Part Five asks the student to discriminate between statements which are pure guesses, statements which are definite assertions but not backed by evidence, and assertions backed by evidence. None of these parts were used in our evaluation for a variety of reasons, the major ones being that the items overlapped Watson-Glaser tests, were too heterogeneous, and seemed to measure general reasoning ability rather than a quality of thought that was subject to the kind of training provided in our program. We did, however, select Parts Six and Seven for inclusion in the evaluation. Both of these parts required the student to make a careful analysis of a persuasive message. Part Six, after presenting a number of arguments against government health insurance, asks such questions as:

1. Which of these is the main conclusion the writer would apparently wish his readers to reach?

2. Which of these thoughts or statements of the writer, even if true, has the least relation to his main argument?
3. Which of these is an assumption made by the writer which is basic to his argument?

These questions are followed by a number of statements which the student must classify into one of the above categories.

Part Seven, after presenting an advertisement for "Roll-Away Windows," asks the student to categorize a series of statements as one of the following:

1. The writer states this, although maybe not in just these words.
2. The writer does not state this, but he does state something which shows that he took it for granted.
3. The writer does not state this, nor does it have any relation to his argument.
4. The writer does not state this, and it would weaken his argument if he did state it.

This part of the test appears to measure how well the student understands the original message, as well as his ability to sort out statements as more or less relevant to the argument.

If time had been available, and the students could have tolerated any more testing, we would have given the complete Watson-Glaser and Michigan State Form A Batteries. Because of these pressures, however, it was necessary to be selective. With so little information on the validity of these two tests (both are fairly reliable), we selected those subjects which seemed most relevant to the experimental curriculum.

2. Iowa Test of Educational Development No. 5, Interpretation -- Social Studies. This test is essentially a critical reading test. It consists of a series of passages to be read, and a series of questions based on each passage. The following questions are illustrative:

What is the passage about?

What is the writer's main purpose in this passage?

With which of the following statements would one expect the writer to be in greatest agreement?

The student is to select the best of four possible choices given after each question.

Measures of Problem Analysis: Experimental Tests*

1. SIAT No. 1: Argument Analysis Test: AAT is a group administered pencil-and-paper test with multiple choice responses. The purpose of the

*developed by Project

measure is to assess how well the student can identify the intellectual operations in an argumentative dialogue. The following type of items follow an argumentative dialogue:

6. What does Mike's statement (f) do in the discussion?

- ☐ (a) Suggests that a second value is involved
- ☐ (b) Suggests that a source is unreliable or untrustworthy
- ☐ (c) Asks for clarification of a word or phrase
- ☐ (d) Takes a more general position in the argument
- ☐ (e) Shows that Joe's position is indefensible

2. SIAT No. 2: Argument Description and Rebuttal Analysis (ADRA): Like AAT ADRA is a group administered pencil-and-paper test with multiple choice responses. The purpose of the test is to assess how well the student can describe the substance of an argumentative dialogue (rather than analyze the intellectual operations), and whether or not the student can select the best rebuttal statements which might be used to counter statements made in the initial dialogue in order to move the discussion along. Since about half the test requires this type of rebuttal analysis, an illustrative item is given below.

(An initial statement is given quoted from the original dialogue.)

13. Now the unions are led by a group of sharp gangsters who are after the worker's pocketbook.

(The student is asked: Which two of the following five statements constitute the best rebuttals to clarify the argument or move the discussion along?)

- a. Do you think all unions are led by dishonest leaders or do you have some particular unions in mind?
- b. You're defining gangster as anyone who wants to give the underdog a fair chance.
- c. Would you clarify exactly what you mean by worker's pocketbook?
- d. What evidence do you have that the worker's pocketbook is at all affected by how much money the companies make?
- e. Do you mean that union leaders are actually breaking laws?

3. SIAT No. 3: Argument Analysis Test -- Open-Ended: This test attempts to get at the same type of analytic competence as the AAT above, but is administered to individual students orally. The student is given a dialogue to evaluate. He is asked:

- A. What are the main points of disagreement in the argument?
- B. How would you go about clarifying or settling these points of disagreement?

The student's oral responses are then coded according to whether or not he correctly identifies points of disagreement, how explicitly and at what level of generality he is able to identify the disagreements, the number of appropriate strategies the student can identify as ways of handling the disagreement, and the number of analytic concepts used in describing the strategies for handling the disagreements.

4. SIAT No. 4: Socratic Discussion Test -- (SDT): The behavior evaluated in this test is either an oral defense of a position in the face of Socratic questioning by an interviewer or behavior elicited in a free response discussion. The student responses are then scored from tapes by a system of content analysis. The categories of behavior for which the student is given credit are as follows:

- a. Use of value judgments and legal claims phrased in general ethical or legal terms
- b. Use of specific facts and sources
- c. Explicit concern with definitional problems
- d. Use of comparative cases and analogies
- e. Explicit concern with relevance of statements
- f. Use of selected dialectic operations:
 - 1. giving a specific statement to support or clarify a general statement
 - 2. generalizing correctly from specific statements
 - 3. pointing out specific contradictory information
 - 4. qualifying a general statement
- g. Use of general analytic statements to conceptualize argument at a more abstract level.

Tests Measuring Knowledge of Social Studies Content

Any experimental curriculum which necessitates the displacement or condensation of social studies content will be criticized by those committed to the present curriculum. One can meet this criticism in at least two ways. The first defense can be made in general ethical and pragmatic terms. This is basically the function of Part One of this volume: to argue that the proposed change in curriculum will better meet the needs of the individual and the community. This also includes the straightforward logical argument that, in fact, the experimental curriculum is actually more directly derived from common teaching objectives than is the traditional curriculum. A second defense becomes pertinent when the criticism raises an empirical question about the relative effectiveness of the two curricula in achieving the objectives of standard curricula, such as the teaching of basic historical information, knowledge about government, geographical skills and facts, etc. It is useful to gather data in order to shed light on such criticism. We, therefore, included in our assessment battery three tests which measure the type of information given high priority by standard curricula. All three tests are standardized measures.

1. Iowa Test of Educational Development No. 1, Understanding Basic Social Concepts. This is a comprehensive social studies test including questions on history, geography, economics, and government. The items are of the standard multiple choice type. For example:

What is a trial jury?

- 1) A court decision
- 2) A crime for which a man is tried
- 3) A promise to tell the truth and the truth only
- 4) A body of people who hear evidence and are expected to bring a verdict

2. The Principles of American Citizenship Test, Form A. This test is similar in form to ITED No. 1. Rather than being a comprehensive test of social concepts, however, almost all items are related to concepts or information about American government. It was developed by the Columbia University Teachers College Citizenship Education Project. The following item is illustrative:

Senator Smith has been talking on the Senate floor for several hours without rest. He is against a bill before the Senate and he is trying to delay or prevent its passage by the method of

- M. filibustering
- N. picketing
- O. logrolling
- P. pork barreling

3. California Tests in Social and Related Sciences, Parts One and Two. This is a comprehensive test of American history information intended for grades nine through twelve. Parts One and Two each take 45 minutes to administer.

A Measure of Interest in Public Issues

We were, of course, interested in the extent to which the experimental curriculum increased or depressed the student's interest in public affairs. Measuring interest is a difficult matter mainly because of the student's desire to please (or displease) the experimenter or teacher. For this reason, we rejected simple questionnaire methods in which the student might be asked "How well did you like the unit on?" or "How did you like the teaching in the last two years compared with the way you studied social studies before?"

Instead we used an instrument employed with some success in earlier studies and very similar to an instrument developed by Columbia's Citizenship Education Project. It has been "successful" in the sense that it tends to reflect treatment differences. The test is called the Newspaper Headlines Test and asks the student to rank sets of newspaper headlines according to which heading describes an article the student would be most interested in

reading and the one he would be least interested in reading. Since there are only three headlines to a set, this results in ranking each set in the order of the student's personal interest.

The directions given to the student are essentially as follows:

There are thirty sets of headlines. For each set put on your answer sheet a plus (+) beside the letter of the headline that seems most interesting to you, a minus (-) beside the headline that seems least interesting, and leave the line beside the letter of the third headline blank.

Following is an example of a set:

- a. New Church Dedicated on Park Avenue
- b. Many Miners Jobless, Families Suffer
- c. Four French Dressmakers Design Original Styles

The test contains three groups of sets. One group contains one headline in each set relevant to a topic included in the experimental curriculum, e.g., "Many Miners Jobless, Families Suffer." A second type of set contains one headline relevant to a social problem not included in the experimental curriculum, e.g., "Group Calls for Ending Sale of Alcoholic Drinks." This group was included to determine whether any increase in student interest in topics studied would also be reflected in an increase in interest in social problems in general. A third group of dummy sets contained headlines relevant to no general social problem. These were included to mask the fact that certain topical areas recurred in the headline sets.

While we recognize the difficulties inherent in this kind of test, e.g., the question of whether or not a student would make the choice in a non-school situation, it is the best interest measure we could devise within the resources at our disposal. Our major effort, as should be obvious at this point, was in the direction of measuring intellectual skills rather than affective learning outcomes.

Summary of Testing Program

1. Standard Tests of Analytic Thinking

Wagmis was administered in the fall and spring of 1959 and 1960 during the two years of experimental teaching to the experimental group and two control groups (West and Wheat). A third control group came into the study late and took the test three times, missing the first fall administration.

ITED No. 5 was administered exactly the same number of times and to the same group as Wagmis.

2. Experimental Tests of Analytic Thinking

SIAT No. 1, the Argument Analysis Test, was developed late in the Project and was administered to the experimental group and to a sample of students in one control school (Wheat).

SIAT No. 2, Argument Description and Rebuttal Analysis Test, was developed before the experiment actually began and was administered in the fall of 1959 and the spring of 1961 to the experimental group and two control schools (West and Wheat).

SIAT No. 3, Argument Analysis Test, Open-Ended, was developed during the first year of experimental teaching. A trial version was administered to the experimental group and one control school in the fall of 1960. A more refined version was administered in the spring of 1961 to the same groups. The initial experimental version was used as a covariance adjustment variable.

SIAT No. 4, Socratic Discussion Test, was developed at the same time as SIAT No. 3 and administered to the same groups at the same time. The response coding system was also used to analyze group discussions of experimental students taped in the spring of 1960, the fall of 1960, and the spring of 1961. Similar discussions were taped and analyzed for one control school (in the fall of 1960 and the spring of 1961).

3. Tests of Social Studies Content

ITED No. 1 and the Principles of American Citizenship Test were administered to the experimental group and two control groups (West and Wheat) in the fall of 1959 and the spring of 1961.

The California Tests in Social and Related Sciences, Parts One and Two, (American history) were administered twice, in the spring of 1960 and the spring of 1961. The first administration was given to three types of groups all of whom were students within the Concord school system. (a) Eighth grade students who had just completed the experimental program; (b) Eighth grade students who had just completed a standard program of chronological American history and geography; and (c) Ninth grade students who had completed a standard course in American history and geography one year earlier. The second administration was given to two types of groups: (a) Ninth grade students who had had the experimental program in grades seven and eight; and (b) Ninth grade students who had had the standard American history course in grades seven and eight. These groups of students were the same as eighth grade groups (a) and (b).

.. Interest in Public Issues

The Headlines Test was administered in the fall of 1959 and the spring of 1961 to the experimental group and two control groups (West and Wheat).

Results

Statistical Treatment of Data

In estimating the effect of the experimental curriculum on learning outcomes, we used the analysis of covariance technique. The purpose of this statistical method is briefly described by Edwards:

The method of analysis of experiments in which adjustments are made in the data for the experimental variable, in terms of data collected on another variable which may condition the outcomes of the experiment, is known as the analysis of covariance. The analysis of covariance is a synthesis of the methods of regression and the methods of analysis of variance. The analysis of covariance is applicable to any experiment in which a source of variation, which it may not be possible to equalize between the various experimental groups prior to the experiment proper, can be measured. An adjustment is then made for this source of variation in the analysis of the outcomes of the experiment.³⁸

Our reasons for using analysis of covariance are two-fold, but both of them fall within the general description provided by Edwards. In the first place, we were interested in comparing the learning increments of the students in the experimental groups with those of the control groups. A possible procedure for making these comparisons is to acquire gain scores by subtracting each individual's pre-test score from his post-test score, and then to determine whether the mean gains for the different groups are significantly different by using the t-test, or with multiple groups, the analysis of variance. However, this procedure assumes that the pre-test and post-test used to obtain each gain score are comparable. This assumption is not tenable with our SIAT No. 3 and No. 4, because we changed scoring procedures between the time of scoring the pre-test administration and the post-test administration. There is, moreover, enough variation in conditions of test-taking to raise a legitimate question as to whether a pre-test and post-test are ever completely comparable even when the same test or an alternate form is used. Therefore, rather than comparing mean gain scores, it is more defensible to consider the students' ability to perform the specialized intellectual operations required on the tests, as measured by pre-test performance, as a source of variation which might not be equalized, and then to adjust the post-test scores for this source of variation. In essence, what this does is allow the researcher to estimate the mean of post-test scores that would have been obtained on a test had the groups been comparable with respect to the uncontrolled variable, i.e., the pre-test scores on the same test.³⁹ It should be noted also that to the

³⁸Allen L. Edwards, Experimental Design in Psychological Research (New York: Rinehart & Company), p. 335.

³⁹Quinn McNemar, Psychological Statistics (2nd edition, New York: John Wiley & Sons, 1955), p. 347.

extent the learning outcomes measures are not stable, i.e., lack reliability, analysis of covariance compensates for this instability much more adequately than does a comparison of mean gains.

The second reason for using analysis of covariance is perhaps more obvious. In attempting to select control groups which would be roughly comparable to the experimental group in general intellectual ability, we had to operate within the organizational structure of the schools -- particularly in regard to groupings of students for instruction. This meant that although we were able to secure estimates of general intellectual ability in the form of intelligence tests, we could not explicitly match groups on this variable. Analysis of covariance, to the extent that intelligence was related to test performance and the groups actually differed on measures of intelligence, allowed us to adjust our post-test scores to make our groups comparable with respect to intelligence test scores at the same time that we adjusted for pre-test performance.

Standard Measures of Analytic Thinking

Wagmis Test. The data summarizing gains made on the Wagmis Test over the two year period are presented in Tables 11.6 and 11.7. Two findings turn up here which run consistently through most of the comparisons. The pre-test scores of one control school, Wheat, are higher than the experimental school; and the pre-test scores of another control school, West, are lower. Strangely enough, intelligence test scores for both of these control schools are somewhat higher than those at Concord.

Regarding the central hypothesis, it is clear from both tables that there is a significant difference among the means of the three schools in the one case and the four schools in the other on the post-test scores. The difference is the result, however, not of higher scores for students in Concord but rather of the lower scores for West. The adjusted means among Concord, Wheat, and Big are strikingly similar, especially in Table 11.6. Nor is the difference between the Concord and Wheat means significant at the .05 level in the comparison presented in Table 11.6. This difference is probably the result of sampling error.

From these data one may conclude that the specialized experimental curriculum probably did not affect the Concord student's performance on a standard test of critical thinking. This is not surprising. For if one looks over the experimental curriculum, there is very little emphasis upon the specific types of intellectual exercises required to do well in Wagmis. The major part of the EC is related to ethical-legal analysis and the evaluation of empirical evidence. Evidence from a recent factor analysis of Watson-Glaser and another form of the Michigan State Test (the ACE test), similar to Form A, indicates

TABLE 11.6

COMPARISON OF CONCORD EXPERIMENTAL GROUP WITH
CONTROL SCHOOLS ON WAGMIS TEST WITH MEANS ADJUSTED FOR
PRE-TEST SCORE (FALL, 1959) AND INTELLIGENCE

School	N	Pre-test		I.Q.		Post-test*		Adjusted Mean
		Mean	S.D.	Mean	S.D.	Mean	S.D.	
Concord	109	36.50	7.96	113.51	12.31	44.94	9.56	46.36
West	126	36.70	7.29	118.47	14.48	43.34	7.82	43.28
Wheat	112	40.56	7.07	117.44	11.09	46.74	7.90	45.17

Test of Differences Between Means

	Degrees of Freedom	F-Ratio	Probability Level
Total	2/342	7.61	.001 .01
Concord v. Wheat	1/216	2.25	.05

*All post-test scores, unless otherwise specified, are for tests administered in the spring of 1961.

TABLE 11.7

COMPARISON OF CONCORD EXPERIMENTAL GROUP WITH
CONTROL SCHOOLS ON WAGNIS TEST WITH MEANS ADJUSTED FOR
PRE-TEST SCORE (SPRING, 1960) AND INTELLIGENCE

School	N	Pre-test		I.Q.		Post-test*		Adjusted Mean
		Mean	S.D.	Mean	S.D.	Mean	S.D.	
Concord	108	39.67	9.84	113.56	12.36	44.88	9.58	45.51
West	126	40.18	7.82	118.55	14.61	43.55	7.94	42.72
Wheat	177	41.90	8.21	116.37	11.12	46.47	7.94	45.39
Big	122	39.52	7.58	109.72	10.83	43.74	9.19	45.55

Test of Differences Between Means

Degrees of Freedom	F-Ratio	Probability Level
3/530	6.97	.001

*Spring, 1961

that the major loading of the parts considered here is on a general reasoning factor.⁴⁰ The results of testing here indicate that our curriculum did not affect any such general reasoning component. The one problem with this interpretation is that it does not explain Glaser's initial finding; namely, that a relatively brief treatment did significantly affect scores on an early form of the Watson-Glaser Test. It is likely that the important difference between Glaser's work and our own work is that we did not teach directly for the specific skills required to do well on the Watson-Glaser Test to the extent that Glaser did. Smith's work,⁴¹ which will be described in detail in Chapter Twelve bears out our finding. This is essentially that when one teaches from a problem orientation in which the skills of problem analysis are directed at specific types of problems (in our work, political problems), rather than at critical thinking skills in general, tests such as Watson-Glaser, Michigan State, Form A, and ACE are not sensitive to the learning outcomes which result. In short, these standard tests of critical thinking probably are sensitive to two kinds of curricular treatments: (a) those which might affect general reasoning ability; and (b) those which might teach the student to handle the specific kinds of reasoning exercises contained in the tests. The negative findings in our own experiment would indicate that the treatment did not have a sufficiently profound impact to affect general reasoning, and was not sufficiently geared to the specific exercises in the tests to have a significant effect on the scores of the experimental group.

ITED No. 5. The results of comparing the experimental school and control schools on ITED No. 5 are similar, if slightly more ambiguous, than those obtained with Wagmis. There appears to be a rather consistent order of schools: West, Big, Concord, and Wheat. While the difference indicated in Table 11.8 between Concord and Wheat is not significant, the difference between Concord and West is. Given the trend established in our discussion of the Wagmis data, plus the relative position of Wheat, it is more reasonable to conclude that there is something deficient in West than it is to conclude that the experimental curriculum had a special effect in Concord. It is also interesting to note the consistently higher mean scores on these two reasoning tests in Wheat. The Wheat mean pre-test scores are higher than the other three schools and this order is maintained, even when the post-test means are adjusted for pre-test scores. Wheat students may be developing in reasoning ability of the type required to do well on this test at a faster rate than the

⁴⁰Velma I. Rust, R. Stewart Jones, and Henry F. Kaiser, "A Factor-Analytic Study of Critical Thinking," The Journal of Educational Research, LV (March, 1962), pp. 253-259.

⁴¹B. Othanel Smith and others, "A Study of the Logic of Teaching: The Logical Structure of Teaching and the Development of Critical Thinking, a Report on the First Phase of a Five Year Project" (Prepared at the University of Illinois, Urbana under Cooperative Research Project No. 258 (7257) of U. S. Office of Education, 1952 [Mimeographed]).

TABLE 11.8

COMPARISON OF CONCORD EXPERIMENTAL GROUP WITH
CONTROL SCHOOLS ON ITED No. 5 WITH MEANS ADJUSTED FOR
PRE-TEST SCORE (FALL, 1959) AND INTELLIGENCE

School	N	Pre-test		I.Q.		Post-test		Adjusted Mean
		Mean	S.D.	Mean	S.D.	Mean	S.D.	
Concord	107	22.56	7.15	113.85	11.73	32.83	10.28	33.98
West	129	20.50	7.17	118.53	14.69	31.20	9.73	31.53
Wheat	154	24.33	8.03	116.87	10.80	36.10	9.29	35.20

Test of Differences Between Means

	Degrees of Freedom	F-Ratio	Probability Level
Total	2/415	11.99	●.001
Concord v. Wheat	1/286	2.09	●.05
Concord v. West	1/231	5.11	●.05 >.01

TABLE 11.2

COMPARISON OF CONCORD EXPERIMENTAL GROUP WITH
CONTROL SCHOOLS ON ITED No. 5 WITH MEANS ADJUSTED FOR
PRE-TEST SCORE (SPRING, 1960) AND INTELLIGENCE

School	N	Pre-test		I.Q.		Post-test		Adjusted Mean
		Mean	S.D.	Mean	S.D.	Mean	S.D.	
Concord	107	26.52	9.76	113.34	12.35	32.75	10.50	32.79
West	134	24.95	10.44	118.69	14.58	31.24	9.75	30.53
West	192	28.02	10.15	115.69	10.88	35.90	9.39	34.11
Big	126	22.30	7.79	109.80	10.56	28.54	10.95	31.97

Test of Differences Between Means

Degrees of Freedom	P-Ratio	Probability Level
3/554	8.72	0.001

other students, although the rate of growth was not quite fast enough in comparison to lead to significant differences within the two year experimental period. It is clear that the Wheat students scored higher, on the average, when the experiment began, and that they maintained this position at the end of the experimental period.

Experimental Tests of Political Analysis

The four varieties of Social Issues Analysis Tests (SIAT), designed specifically for the Project, have been discussed in detail. It should be noted, before proceeding to the presentation of data, that these tests do not demand a technical vocabulary, even though to do well the student must conceptualize a political argument in other than the specific substantive terms of the controversy. Our evidence indicates that performing well on the SIAT does not require the technical competence available only to specially trained personnel. The pencil-and-paper tests, i.e., SIAT 1 and 2, were administered to graduate students in social studies education. These people who had no special training, aside from a general liberal arts education, performed extremely well on the tests according to our criteria of correct responses.

A related question is relevant to our earlier discussion of Glaser's research. That is, could any gains which favored the experimental students be due to our teaching specifically for our own measures? This question, which is often asked of experimenters who attempt to evaluate experimental curricula, reveals a basic misunderstanding of the relationships among objectives, instructional methods, and evaluation in the process of curriculum development. An adequate answer to what is at best an inadequate question must begin not with the tests but with the basic problem of curricular design and evaluation. Fundamental knowledge of the factors affecting "transfer of learning" dictate that the instructional context should be as similar to the application situation as is possible. Evaluation, to be valid, moreover, should also take place in a context as similar as possible to the one in which we predict the "ultimate" criterion behavior is most likely to occur. It follows, then, that to the extent that these two conditions are fulfilled, the instructional and evaluational contexts will be similar because of their relation to the objectives of instruction. One will be "teaching for the test" because of its logical relationship to the process of curriculum development. To the extent that the teacher or researcher is interested in transfer of the training which occurs as a result of the initial instructional situation, a variety of test situations may be used to evaluate the generality of the learnings.

Perhaps we should now point out that the procedure we followed is in line with the above considerations. We began with a conceptual approach to political controversy and then developed materials and procedures to teach this approach. At the same time, it was recognized that if the experimental curriculum were to be adequately evaluated, tasks had to be devised which could reflect, within appropriate settings, increments in the particular skills which were sought. Thus, a considerable effort was made to devise means of assessing whether or not the students had learned to handle the conceptual scheme within the context of political controversy. To the extent that the tests involved

the components of the system to be taught, and to the extent the situational contexts were similar for instruction and evaluation as dictated by our objectives, it could be said, in one usage of the phrase, that we taught for the tests. This distorts, however, the very legitimate relationships which should exist between objectives, instruction, and testing pointed out above.

It is also worth noting that the complexity of the conceptual scheme presents a challenge both to the teacher and the student. But the complexity of the analytic framework is coupled in the teaching situation with the substantive complexities of particular political issues, providing a task which is difficult and taxing. There is no easy assurance that the educational experiences of the classroom will transfer to other controversial situations.

SIAT No. 1, Argument Analysis Test. As we indicated earlier, SIAT No. 1 was developed late in the Project, and therefore we have no pre-test measure. For this reason we were able to administer it to a limited sample in only one control group (Wheat). The results are presented in Table 11.10. One should note not only the means but the standard deviations. The test scores clearly reflect the results of experimental training. It is probably the most direct measure of the student's ability to understand the conceptual system. (This does not necessarily mean, however, that the student can relate the conceptual system to an argument which he himself is attempting to devise or rebut.) Although the control sample is limited, it is probably safe to conclude from these data that this test is sensitive to the kind of training given the experimental group. It is very important to qualify this finding, however, in terms of the broader objectives of the Project. We were interested, in the final analysis, in the student's ability to carry on a more intelligent dialogue when he discussed contemporary political issues. This test does not pretend to measure any such ability. It tells us to some degree whether the student has the conceptual tools to analyze a dialogue (in which he might be a participant); it does not tell us whether or not he has the ability or is inclined to actually apply the tools to a dialogue in which he is a participant. Even with this qualification, however, the results indicated in Table 11.10 are encouraging. It means that the student at this age and grade level has the intellectual capacity to be taught to think in these abstract conceptual terms.

SIAT No. 2, Argument Description and Rebuttal Analysis. Four forms of the SIAT No. 2 tests were developed, each based on a dialogue focusing on a different topic, but otherwise identical in form. The four dialogues related to (a) the honesty and responsibility of labor leaders, (b) the equal treatment of Negroes in the South under segregated conditions, (c) the possible dangers of atomic fallout as against the testing of atomic bombs for national defense, and (d) the quality of public education. In abbreviated form the four topics were designated labor, integration, fallout, and education. These particular topics were selected for two reasons. First, all were regarded as sufficiently permanent controversial issues that the tests would not become dated over the two years of the experimental treatment. Secondly, two topics were deliberately chosen because they were included in the experimental

TABLE 11.10

COMPARISON OF CONCORD EXPERIMENTAL GROUP WITH A CONTROL SCHOOL
ON SIAT No. 1 WITH MEANS ADJUSTED FOR INTELLIGENCE

School	N	Post-test		Adjusted Mean
		Mean	S.D.	
Concord	109	12.36	4.40	12.79
Wheat	33	10.73	2.63	9.37

Test of Differences Between Means

Degrees of Freedom	F-Ratio	Probability Level
1/138	24.52	.001

curriculum: labor and integration; and two topics were chosen because they would be deliberately excluded from the experimental curriculum: fallout and education. In this way we hoped to test whether or not various subtests would be affected by the specific topics included in the experimental curriculum, or whether the student was learning a general set of analytic skills which could be applied to political issues similar to those studied. The evidence in Table 11.11 indicates that the latter is the case. Apparently there was no special benefit from having studied a topic in the experimental curriculum, which happened to be the basis of the dialogue in two of the subtests. The difference between the post-test means of the two sets of subtests is probably due to measurement error -- (the value of t is only 1.26)--although the direction of the difference is right if one were to assume that studying a topic is of direct help on a subtest.

We can now ask the more important question. Assuming that Concord students could make the gains suggested in Table 11.11 on any subtest of the SIAT No. 2, are these gains significantly greater than students would make simply on the basis of maturity and having studied a conventional history and/or geography curriculum? The findings relevant to this question are presented in Table 11.12. The evidence is fairly clear that SIAT No. 2 is sensitive to the effect of the experimental curriculum. This finding is substantially more important than the difference between experimental and control groups on SIAT No. 1, for here the student is not only required to conceptualize the argument in more general terms, but he must also relate specific abstract descriptions of the argument to specific rebuttal statements made to meet the challenge of a prior statement. In other words, here the student is evaluating new substantive statements made in response to statements in the initial dialogue. Moreover, he is given no conceptual analytic cues by which this evaluation can be carried out, for all the choices are in terms of the substance of the disagreement. To do well in the test the student must have either an intuitive or implicit sense of political analysis, or he must use the conceptual framework provided him in the experimental curriculum to evaluate what is actually happening in the argument. It seems clear that the fact that the Wheat students do substantially better than the West students may indicate that factors such as the extent to which nonsystematic training occurs in the school or the home may affect students not explicitly trained in political analysis. For some reason, the Wheat students appear to have an advantage over the West students, although the curriculum of neither school provides such systematic training. What is more significant is that apparently this kind of competency can be taught to students at this level of maturity, as indicated by the significant increase in the experimental students' scores.

Again, however, a note of caution. This does not mean that the students who studied the experimental curriculum can necessarily carry on a more analytic or intelligent political dialogue. It means only that he can better evaluate statements made by others in a dialogue. This, of course, is an important step in the right direction, but it is not the final objective of our work.

TABLE 11.11

MEANS OF CONCORD STUDENTS ON SIAT No. 2, SUBTESTS BASED ON TOPICS
NOT STUDIED IN EXPERIMENTAL CURRICULUM

N	Pre-test Means (fall 1959)		Post-test Means (spring 1961)	
	L & I*	F & E	L & I*	F & E
107	19.97	19.07	23.21	23.64

*Indicates subtests in which topics were included in experimental curriculum

TABLE 11.12

COMPARISON OF CONCORD EXPERIMENTAL GROUP WITH
CONTROL SCHOOLS ON SIAT No. 2* WITH MEANS ADJUSTED FOR
PRE-TEST SCORE** AND INTELLIGENCE

School	N	Pre-test		I.Q.		Post-test		Adjusted Mean
		Mean	S.D.	Mean	S.D.	Mean	S.D.	
Concord	109	18.79	5.05	113.51	12.31	23.68	5.08	24.28
Vest	115	19.61	4.01	117.50	14.82	20.74	4.36	20.36
Wheat	173	20.20	4.61	116.61	10.88	22.28	4.32	22.15

Test of Differences Between Means

	Degree of Freedom	F-Ratio	Probability Level
Total	2/392	26.70	.001
Concord v. Wheat	1/277	19.54	.001

*Combined forms Labor and Integration

**Henceforth, all pre-test scores were obtained in the fall of 1959.

SIAT No. 3, Argument Analysis Test: Open-Ended. The open-ended version of the argument analysis test was an effort to see to what extent the student could conceptualize an argument at the more abstract level demanded by the exclusion of cues provided by precoded multiple choice test. It was administered individually to students and their taped responses were later coded. To do well on the test requires essentially the same skills as those indicated for SIAT No. 1, except that the scoring was liberalized so that the test probably penalizes the student less for not having specialized or technical knowledge in argument analysis.

Unfortunately the resources of the Project did not allow us to administer the test in more than one control school. This is the more regrettable because the school in which we were able to give the test (West) did consistently worse on other tests than did both the experimental school and the other main control school (Wheat). As indicated on Table 11.13, however, the differences between Concord and West are so striking on this particular test that it is probable that the Concord scores would also have been significantly higher than Wheat. There is also some indirect evidence in the fact that Wheat students scored significantly lower than Concord students on SIAT No. 1, which is a similar test.

One methodological note of caution should be given here. Four people did all of the interviewing in both the experimental and control schools. Three of the interviewers were also experimental teachers. In no case, however, did a teacher interview his own students. All of the teachers were, however, known to be associated with the Project. The fourth interviewer was not known by experimental students, and was not known to be associated with the Project in any direct way. In checking out the performance of the "non-Project" interviewer, moreover, we found no difference suggested between the way his interviewees performed and the way the other students performed. There appeared, in fact, to be no systematic difference among interviewers. It should also be noted that all of the interviewers were experienced teachers and none reported any problem developing rapport with the students in either the control or experimental school.

SIAT No. 4, Socratic Discussion Test: Interview Situation. The SDT is probably the most direct approach to measuring the objectives of the Project. In it the student was involved in a defense of a policy stand with respect to a specific case under discussion. The central difficulty with the test is, of course, the lack of standardization of the test situation. Each interviewer or socratic questioner had before him a list of the major arguments which might arise, and strategies for handling these arguments. In a sense, each was armed with the same arsenal from which he might select points to challenge the students defense. But the questions he actually asked and the points he actually made were conditioned by the argumentative tactics made by the student. Thus, some students may choose to deal only with the factual or legal claims involved in the case; others may choose to deal with the ethical problems. Each interviewer was instructed to deal with as many issues as could naturally be inducted into the discussion, but not to effect abrupt transitions or totally rechannel the structure of thought. This, we thought,

TABLE 11.13

COMPARISON OF CONCORD EXPERIMENTAL GROUP WITH A CONTROL SCHOOL
ON SIAT No. 3 WITH MEANS ADJUSTED FOR PRE-TEST SCORE AND INTELLIGENCE

School	N	Pre-test*		I.Q.		Post-test		Adjusted Mean
		Mean	S.D.	Mean	S.D.	Mean	S.D.	
Concord	106	18.02	7.03	113.43	12.14	30.76	10.81	31.15
West	59	17.08	6.37	120.66	14.86	19.19	6.68	18.49

Test of Differences Between Means

Degrees of Freedom	F-Ratio	Probability Level
1/160	73.39	.001

*A somewhat different scoring system was used on the pre-test than on the post-test, but presumably both were measuring the same attribute. These numbers, however, cannot be considered precisely comparable to the post-test numbers. Covariance, of course, does not make the assumption of precise comparability.

Donald W. Oliver, "The Experimental Process in the Study of a Social Problem" (Unpublished doctoral thesis, Harvard Graduate School of Education, 1955).

would have added an additional unwanted dimension to the test situation. Before we discuss the results, we should also point out that the SDT is a highly experimental measure which, although potentially the most powerful evaluation instrument developed by the Project, is still undergoing modification and study to improve its reliability and establish its validity.

Like SIAT No. 3, this test was administered in only one control school (West). As we pointed out above, it would have been better to use Wheat, but our limited resources precluded this possibility. Table 11.14 indicates that the measure is sensitive to what is apparently a relatively small but systematic difference between the control and the experimental group in favor of the Concord group. The reader should be warned, however, that the pre-test and the post-test are not comparable forms. The pre-test was shorter and, therefore, the difference between pre-test and post-test means cannot be considered "gain" scores in any sense of the word. One other fact might be noted here which helps in the interpretation of these data. Although the experiment began in the fall of 1959, the pre-test was not developed and actually administered until the fall of 1960. The Concord students thus had undergone some fairly intensive treatment before the administration of the pre-test. This might very well account in part for the substantial difference between Concord and West on pre-test means (1.62 in favor of Concord). This fact, of course, makes it more difficult for Concord students to show a significantly higher score on the post-test than the West students. This interpretation is further supported by data presented below in which the SDT was applied to behavior obtained in peer group discussions, rather than in interviewer-student discussions.

Use of the Socratic Discussion Test for the Analysis of Student Led Discussions

In addition to the regular testing situations described above, we had decided from the outset to use a somewhat unique but realistic setting to assess the student's ability to frame a political problem and handle it analytically. This is the student led group. That high school students are able to handle such a situation had been established by Oliver.⁴² The setting for this discussion has already been described in Chapter Ten. To recapitulate, approximately 12 students, seated in a semicircle, are given a controversial case. The case is read aloud to the group by an adult, and each member of the group follows on his own copy. The group is then asked to reach common agreement on how the problem in the case should be handled, and the adult administering the situation leaves the room. All of these situations in our work have been recorded, which gives the student added incentive to "perform" well. (Junior high school students in our experience are NOT "mike shy.")

⁴²Donald W. Oliver, "The Discussion Process in the Study of a Societal Problem" (Unpublished doctoral thesis, Harvard Graduate School of Education, 1956).

TABLE 11.14

COMPARISON OF CONCORD EXPERIMENTAL GROUP WITH A CONTROL SCHOOL
ON SIAT No. 4 WITH MEANS ADJUSTED FOR PRE-TEST SCORE AND INTELLIGENCE

School	N	Pre-test		I.Q.		Post-test		Adjusted Mean
		Mean	S.D.	Mean	S.D.	Mean	S.D.	
Concord	103	10.26	3.81	113.71	12.19	19.53	10.11	19.39
West	59	8.54	4.23	120.66	14.86	15.45	10.54	15.72

Test of Differences Between Means

Degrees of Freedom	F-Ratio	Probability Level
1/157	4.28	<.05 >.01

Also review methodological problems should be noted here. After the data were initially scored with the SPSS content analysis instrument, it became evident that there had been rapidly deteriorating reliability on the domain categories. This was not discovered until it was too late to re-score. Consequently, we decided to eliminate those categories from the "valued acts." Initially, it was discovered that groups often made extraordinarily high scores simply by talking about the case. This occurred because we had carried over from earlier work a statement directly from the case was scored as a "specific claim" and any mention of the case was scored as a "source." As stated in Chapter VIII, "specific claims" and "sources" are both valued acts. In retrospect this seemed to be an unwise decision, and we therefore eliminated those two categories from the category "valued acts." Some of these instances were post hoc, occurred only after the data seemed so erratic as to need to be scored at all. We realize that this raises some serious methodological questions, and that our only recourse for a larger confirmation of the hypothesis being tested here is a replication of this work in a similar study.

This kind of situation imposes two very different kinds of problems on the group: procedural problems and task problems. Initially the group must decide who is going to be the leader and how the discussion is to be channeled in an orderly fashion. Some groups, of course, are never able to "solve" this problem, and the discussion turns out to be disorganized and chaotic. Most groups, however, immediately select a chairman and get down to the business of dealing with the case. Only after the initial frustrations of attempting to deal with a highly controversial case are felt do procedural problems again erupt, usually in the form of deliberate attempts at tension release (jokes and wisecracks) or status rivalry of group members (the nominal leader may be challenged by another member of the group). In general, however, it is surprising how well junior high school students handle this type of situation.

In the course of the two year experiment student-led discussion specimens were obtained in February, 1960; May, 1960; September, 1960; February, 1961; April, 1961; and June, 1961. It was impossible to obtain this much discussion time from the major control school (West) because of the many other testing activities that we had carried on there. We ran two sets of discussions there in September, 1960 and June, 1961. Obviously this created a problem of comparable experience with the testing situation: the experimental group obtained a great deal more experience with the testing situation than did the control. The most interesting findings are, therefore, the changes in behavior of the experimental group. These data are summarized below in Table 11.15.⁴³ Only the February, May, June, and September trials are presented because of the scoring problems described in the note below.

The results presented in Table 11.15 suggest the following generalizations. 1) The scores of the experimental group clearly reflect the effects of specialized training. Untrained groups tend to obtain a stable score of about 30 valued acts (excluding dynamic acts). 2) The measure (SDT) as used in this context is sensitive to short intensive teaching of the analytic model.

⁴³Two serious methodological problems should be noted here. After the tapes were initially scored with the SDT content analysis instrument, it became evident that there had been rapidly deteriorating reliability on the dynamic categories. This was not discovered until it was too late to rescore. Consequently, we decided to eliminate these categories from the "valued acts." Secondly, it was discovered that groups often made astonishingly high scores simply by talking about the case. This occurred because of a convention we had carried over from Bales: any statement directly from the case was scored as a "specific claim" and any mention of the case was scored as a "source." As stated in Chapter Ten, "specific claim" and "source" are both valued acts. In retrospect this seemed to be an unwise decision, and we therefore eliminated these two categories from the category "valued acts." Both of these decisions were post hoc, occurred only after the data seemed so erratic as to make no sense at all. We realize that this raises some serious methodological questions, and that our only recourse for a firmer confirmation of the hypotheses being tested here is a replication of this work in a similar study.

This point is clear from the dramatic increase in scores on the May discussions, which followed immediately after a ten day intensive treatment. Ten short cases, modeled after Supreme Court cases, were used specifically to illustrate the process of ethical clarification through the generalization of value statements and the use of analogies. 3) The washout effect caused by the three months summer layoff was much more pronounced for groups composed of students with low or average ability than for the high ability groups. 4) While there was a substantial increment in scores for high ability groups during the second year, they did not surpass the achievement registered at the end of the first year of the program. 5) The groups composed of students with low and average ability never regress to their initial score (28.00), nor did they gain in any significant way.

These findings are extremely enlightening and must be interpreted within the context of the experimental curriculum. As we stated earlier, the first year of the program was mainly concerned with the teaching of analytic skills, but not within the context of any complex societal problem. The unit of analysis was the controversial case, and most of these were relatively brief, serving as an illustrative basis for the concepts we were trying to teach. The climax of the first year came with a two week intensive effort to relate a number of the more important analytic concepts. It is after this point that the high scoring May discussions were held. During the second year an attempt was made to teach the student to place the controversial case within a broader historical context. This is a more difficult intellectual task because the analytic structure of an argument is no longer so schematized; it must encompass a great deal more information. Thus the student's effort to grapple with this additional complexity, and at the same time maintain some conception of the total analytic structure of an argument is probably the reason that intellectual growth during the second year reached no higher point than it had at the end of the first year. And this brings us to a fundamental fact: the measure we were using was not sensitive to an intellectual process that would allow a student to operate within the general framework of the analytic model, and at the same time handle more information at a more sophisticated level. In the development of the content analysis system upon which the SDT is based, we struggled a great deal to differentiate this type of substantive factual and conceptual sophistication within the categories "general claim" and "specific claim," but no reliable method was discovered.

In the light of these facts we would suspect that the able groups in June, 1961 were performing at a substantially higher level than they were in May, 1960, but the SDT simply does not pick this up. Our suspicion would be, however, that this is not the case with the slower groups. We were, in fact, much more optimistic about the slower groups' ability to handle the analytic model after the first year of the experimental program than later in the second year. Nor is this surprising. Providing the student with historical and conceptual background, as we did the second year, simply adds a great deal of "noise" which may well interfere with the direct application of the analytic concepts to the controversial case under consideration. The most fundamental type of understanding of a problem or issue from the point of view of this study would allow the student to go from the substance or concrete facts

TABLE 11.15

COMPARISON OF STUDENT LED GROUPS IN CONCORD AND WEST
SCORED BY MODIFIED SIAT No. 4 CONTENT ANALYSIS SYSTEM

Date of of Trial	West		Concord			
	N*	Mean	N*	Total Mean	High Ability	Low Ability
Feb. 1960			10	31.90	34.50	28.00
May 1960			10	71.70	74.00	68.25
Sept. 1960	12	31.08	19	50.11	59.00	40.57
June 1961	12	25.08	17	59.82	74.22	43.62

*Total number of discussions equals number of groups times number of individual discussions.

...the training given in these areas in conventional junior high school programs. Table 11.15 indicates that one control school (West) and Concord maintained a similar rate of increase, while the other control school (East) fell significantly behind. This is, of course, consistent with previous findings. It is clear that either the educational program or general educational climate of West is of lower quality than for the other two schools, although the academic aptitude of the students is at least as good if not better than that of the other schools' students.

We might assume perhaps, that the two control schools represent a fair range of quality of schools in well-to-do suburban areas. If this is the case, we have demonstrated that the experimental social studies curriculum is of sufficient quality to allow the experimental group to keep pace with the better of the control schools in conventional social studies as often measured by this particular test.

describing the issue to the analytic concepts which might allow them to place it in a broader historical context. At the end of the first year, the groups were capable of doing this, although somewhat mechanically and "stiffly." In the second year, however, they were not only required to do this, but they felt some responsibility to relate the analytic concepts to a broader body of conceptual and historical information contained in a three to six week substantive unit. It was this task which seemed to overwhelm the slower groups. It is difficult to draw any firm conclusions about the performance of the more able groups. Insofar as the SDT measures analytic ability, there is no gain between the spring of 1960 and the spring of 1961. It is our feeling that many of the students, as a result of the second year, were more easily able to integrate the analytic concepts into their own style of thinking, and were able to handle the substance of the argument better as a result of the factual background provided in the total unit. And this, unfortunately, was probably not assessed by the measure.

These findings are related to a question with which we were particularly concerned: How intelligent must a student be to deal with public controversy analytically? Our data suggest that a somewhat mechanistic scheme can be taught to average and slightly below average students without great difficulty, but that these students have particular difficulty integrating the analytic model into their normal train of thinking, and also using it to handle complex situations. We would certainly recommend that if our study were repeated, there should be a gradual increase in the complexity of issues and the amount of background information provided for a given unit. As it was, we jumped from simple cases to complex units, and only the abler groups had the capacity to rebound from the shock.

Tests of Social Studies Content

ITED No. 1, Understanding Basic Social Concepts. ITED No. 1, while covering a variety of concepts in history, geography, and economics is certainly biased heavily in the area of American history and government. It should reflect the training given in these areas in conventional junior high school programs. Table 11.16 indicates that one control school (Wheat) and Concord maintained a similar rate of increase, while the other control school (West) fell significantly behind. This is, of course, consistent with previous findings. It is clear that either the educational program or general educational climate of West is of lower quality than for the other two schools, although the academic aptitude of the students is at least as good if not better than that of the other schools' students.

We might assume perhaps, that the two control schools represent a fair range of quality of schools in well-to-do suburban areas. If this is the case, we have demonstrated that the experimental social studies curriculum is of sufficient quality to allow the experimental group to keep pace with the better of the control schools in conventional social studies learning outcomes as measured by this particular test.

TABLE 11.16

COMPARISON OF CONCORD EXPERIMENTAL GROUP WITH CONTROL SCHOOLS
ON ITED No. 1 WITH MEANS ADJUSTED FOR PRE-TEST SCORE AND INTELLIGENCE

School	N	Pre-test		I.Q.		Post-test		Adjusted Mean
		Mean	S.D.	Mean	S.D.	Mean	S.D.	
Concord	108	24.05	7.10	113.46	12.35	35.71	10.02	37.18
West	127	23.51	6.74	119.04	14.41	32.58	9.12	32.94
Wheat	180	26.91	7.28	116.57	10.50	38.43	9.06	37.30

Test of Differences Between Means

Degrees of Freedom	F-Ratio	Probability Level
2/410	17.53	.001

Principles of American Citizenship Test. We realized that any findings based on ITED No. 1 might very well be criticized because of the limited sampling of what teachers consider important knowledge in the areas of history and government taught in grades seven and eight. And this would especially be the case if we could demonstrate that the requirements of conventional curricula could, in fact, be met within the framework of the experimental curriculum. In an effort to meet this criticism we administered two additional tests. One is a more intensive test in government; the other a more intensive test in history. The more intensive government test was developed by the Teachers College Citizenship Education Project of Columbia and deals both with the principles and structure of government as well as governmental process. The comparisons between control and experimental schools for this test are presented in Table 11.17. These data more or less replicate the findings reported in Table 11.16. When the students are required to demonstrate a more thorough knowledge of the workings of American government, the experimental students demonstrate a rate of growth comparable to the better control school. The striking difference between West when contrasted with Concord and Wheat is again borne out.

The California Tests in Social and Related Sciences, Parts One and Two (American history)

The title of this test is somewhat misleading. It is clearly a very intensive test in American history, and will therefore be referred to as the California American History Test. Since we had exhausted the time available to us in control schools, and wished to do delayed testing, we decided to deal with a smaller number of subjects already available in the Concord school system. The first testing was done in June, 1961 immediately at the conclusion of the seventh-eighth grade sequence for students in both the experimental and conventional American history program. Since there were approximately two classes (about 60 students) then in grade nine who had never been exposed to the work of the Project in the seventh and eighth grades, these students were also tested at this time. (These students constitute "conventional group B"). Both experimental and conventional groups who had originally been tested at the end of the seventh-eighth grade sequence were again tested at the end of grade nine. (Students at Concord ordinarily take either ancient history or civics in grade nine, neither of which would be expected to contribute to their score on the American History Test.) At this point we were interested in two hypotheses:

- a) Would students who studied the experimental curriculum be penalized on a comprehensive intensive history examination for not having studied a conventional American history course?
- b) How much loss is there in knowledge of historical material after one and two years for both conventional and experimental groups?

It should be noted that since the special high ability group in the school was taught the experimental curriculum, it was excluded from this particular comparison because of the possible interaction between the ability level of the

TABLE 11.17

COMPARISON OF CONCORD EXPERIMENTAL GROUP WITH CONTROL SCHOOLS
ON PRINCIPLES OF AMERICAN CITIZENSHIP TEST WITH MEANS ADJUSTED FOR
PRE-TEST SCORE AND INTELLIGENCE

School	N	Pre-test		I.Q.		Post-test		Adjusted Mean
		Mean	S.D.	Mean	S.D.	Mean	S.D.	
Concord	108	24.71	8.93	113.54	12.36	34.23	7.98	34.76
West	128	21.35	5.75	118.51	14.73	28.59	6.59	29.35
Wheat	174	26.48	6.18	116.61	10.55	35.05	6.90	34.17

Test of Differences Between Means

Degrees of Freedom	F-Ratio	Probability Level
2/405	35.74	.001

group and the quality of the learning situation within the group. This along with the fact that we had a very limited sample and could not set up matched groups, made the covariance statistical model absolutely necessary. The following control measures were therefore used: 1) Kuhlman-Anderson Intelligence Test given in grade six as a control for general academic ability; 2) the Stanford Reading Test given in grade seven; 3) the seventh grade social studies grade point average. This was used as an indication of motivation to do well in social studies, which might not be accounted for either in an intelligence or reading test. The comparison of the groups on these measures is given in Table 11.18. The correlations between the control measures and the total test as well as among the control measures themselves is given in Table 11.19.

The results bearing on the two hypotheses listed above are reported in the first three rows of Tables 11.20 and 11.21. These data clearly indicate that students who studied a conventional course in American history gained no special advantage in information retained, as measured by this test, over students who participated in the experimental curriculum.⁴⁴ The differences which occur are, for the most part, accounted for by slight differences in control measures, i.e., differences between the abilities of the groups. The largest variance ratio occurs between the experimental group and conventional group A at the end of grade eight, and is in favor of the experimental group.

It was this finding which led us to a more intensive analysis of the data to determine whether or not classifying items in a number of relevant ways might not explain this difference. The test was therefore further analyzed and broken down into the following categories of items:

- a) items not studied in the experimental curriculum at all, but which are normally included in a conventional American history course;
- b) items treated in a conventional way in the experimental curriculum, i.e., a standard reading and discussion based on the text;
- c) items treated topically in one of the problem units developed especially for the experimental curriculum;
- d) items treated both in a conventional way as well as topically by experimental groups.

This further analysis is given in rows 4-7 in Tables 11.20 and 11.21. The results shed further light on the value of various approaches taken to the

⁴⁴This result is supported by results obtained in a study by Arnold Rothstein, "An Experiment in Developing Critical Thinking through the Teaching of American History," Dissertation Abstracts, XXI, no. 5, 1960.

TABLE 11.18

MEANS AND STANDARD DEVIATIONS OF CONTROL MEASURES USED IN COMPARING STUDENTS TAUGHT BY EXPERIMENTAL CURRICULUM AND CONVENTIONAL AMERICAN HISTORY CURRICULUM ON CALIFORNIA AMERICAN HISTORY TEST

Group	N	K-A I.Q.		Reading Test		G.P.A.	
		Mean	S.D.	Mean	S.D.	Mean	S.D.
Experimental	56	108.68	11.89	59.27	15.92	3.05	0.90
Conventional Group A	31	112.32	11.34	60.90	15.20	3.35	0.61
Conventional Group B*	59	110.54	12.04	59.91	17.11	2.97	0.74

*Conventional Group B was tested only at the end of Grade 9 during the first testing period.

TABLE 11.19

CORRELATIONS OF CONTROL MEASURES USED IN COMPARING STUDENTS TAUGHT BY
EXPERIMENTAL CURRICULUM AND CONVENTIONAL AMERICAN HISTORY TEST

	1	2	3
1. California History Test (parts I and II)			
2. I.Q.	.56		
3. Reading Test Score	.71	.61	
4. G.P.A.	.62	.52	.53

TABLE 11.20

COMPARISON OF CONVENTIONAL AND EXPERIMENTAL GROUPS ON
CALIFORNIA AMERICAN HISTORY TEST: IMMEDIATE TESTING AT END OF
SEVENTH-EIGHTH GRADE SEQUENCE

California History Subtest	No. Items	Experimental Group (N=56)		Conventional Group (N=31)		F-Ratio (DF=1/81)
		Mean	S.D.	Mean	S.D.	
Part I (pre-Civil War)	115	51.52	12.04	54.03	12.88	0.37
Part II (post- Civil War)	120	53.16	13.61	52.03	13.94	1.96
Total I & II	235	104.66	24.14	106.06	24.92	0.34
Items not studied in experimental curriculum	73	27.46	6.41	28.13	7.33	0.00
Items treated only historically in experimental curriculum	91	43.98	10.24	45.94	10.38	1.39
Items treated only topically in experimental curriculum	49	24.32	6.86	23.25	6.45	3.60
Items treated both historically and topically in experimental curriculum	18	8.98 (9.19)*	3.36	7.71 (7.34)*	2.61	12.87+

*adjusted means

+significant at the .001 level

TABLE 11.21

COMPARISON OF CONVENTIONAL AND EXPERIMENTAL GROUPS ON
CALIFORNIA AMERICAN HISTORY TEST: ONE YEAR DELAY AT END OF GRADE NINE

California History Subtest	No. Items	Experimental Group (N=56)		Conventional Group A(N=31)		Conventional Group B(N=51)		F- Ratio*
		Mean	S.D.	Mean	S.D.	Mean	S.D.	
Part I (pre-Civil War)	115	54.20	12.44	54.74	14.21	54.71	14.16	0.91
Part II (post- Civil War)	120	50.82	13.67	54.67	13.65	52.15	16.13	0.23
Total I & II	235	105.02	24.36	109.42	26.43	106.86	28.70	0.30
Items not studied in experimental curriculum	73	26.39	6.84	29.77	7.54			2.70
Items treated only historically in experimental curriculum	91	46.16	10.45	48.48	12.45			0.03
Items treated only topically in experimental curriculum	49	23.91	6.97	23.26	6.41			2.07
Items treated both historically and topically in experimental curriculum	18	8.55 (8.73)*	3.00	7.90 (7.58)*	2.49			4.33**

*The first three F-Ratios have 2/140 Degrees of Freedom; the last four have 1/81 Degrees of Freedom

*adjusted means

**significant at the .05 level

organization of a history curriculum. The topically organized curriculum tends to have an advantage, as indicated in row 6, but the advantage does not reach statistical significance. In those areas in which the topical curriculum is set within a conventional historical framework, the experimental students have a statistically significant advantage, both in immediate and delayed testing.

Regarding the second hypothesis (how much information loss is there on delayed testing), the results are rather astonishing. There is surprisingly little loss for either experimental or conventional groups. It should be noted that there are actually two tests for this hypothesis. First, since most of the information required to do well on Part I was taught in grade seven and the test was first administered at the end of grade eight, the comparison between Parts I and II for the first testing, as well as the comparison of the two administrations of the test are relevant to the delayed testing hypothesis. The overall results on the delayed testing hypothesis are unstable and can probably be accounted for on the basis of sampling error.

The stability of this measure over time is somewhat difficult to account for. It is probable that there is a reasonable percentage of items covering information that is studied a number of times by the student, and has become part of the lore of the culture, at least the school culture. Assuming, for example, that the student could get one quarter of the items right by guessing and another quarter of the items right by the "lore" hypothesis, one might anticipate that a student would get half the items right without a great deal of detailed knowledge of American history. The students in Concord who took the test do, in fact, on the average, get fewer than half of the items correct. The norms given in the manual for the test, however, tend to contradict this reasoning. A total score of 110 yields a percentile of 50 with a norms group consisting largely of eleventh and twelfth grade students, who have completed a senior high school course in U. S. history. The Concord average score is 107 which is very close to the 50th percentile. One might conclude that either the norms are spuriously low or the Concord group is exceptionally able. Our own inclination would be toward the former interpretation.

Regardless of what causes the stability of these scores over time, however, the measure yields three important findings. First, the experimental curriculum, which tended to neglect large areas of history in favor of intensive analysis of a limited set of problems did not prove itself deficient as a conveyor of conventional knowledge about American history. Second, when the student learns information both within the conventional context of a history text as well as within the framework of a particular problem, he tends to remember such information better than when he studies it only within the context of history. And third, there is surprising stability over time in what students have learned about American history by the end of the eighth grade.

Results: Student Interest in Contemporary Issues

As indicated above the Project developed a relatively simply measure of interest in particular categories of problems. The measure is called the

Headlines Test, because it requires the student to rank sets of three headlines in the order in which he asserts he is most likely to read the articles following the headlines. The test actually includes two subtests and a number of dummy items as follows. Subtest P: 11 sets include a headline related to a problem studied intensively by the experimental group. Subtest G: 12 sets include headlines concerning general national or international problems, such as crime, financing of education, and Berlin, which were not studied by the experimental group. Headlines in the remaining sets as well as the other two headlines in critical sets suggested what might be called human interest stories, e.g., "Twenty men saved after 7 hours among sharks," "Children escape falling brick wall by inches," etc. For reference purposes we shall refer to the two measures as dealing with "Project Topics" and "General Topics."

The findings are summarized in Tables 11.22 and 11.23. They are interesting, but very difficult to interpret. It appears that the experimental group gained very little interest in the topics studied as part of the experimental curriculum. On the other hand, the control groups which had not studied these issues, showed significantly greater gain on this part of the test. The fact that the experimental group evidenced little change is not so difficult to explain as the fact that the control group made such striking gains. One might speculate that as adolescents are just beginning to attain their full intellectual power, broader ideological issues begin to become more important. Problems bearing on racial and ethnic issues as well as organized labor constituted six of the 11 test items related to what the Project had studied. These topics are clearly more ideologically oriented than such issues as government corruption, crime, education, and the farm problem. Possibly, then, it is easier to polarize one's view on the race or labor issue, and consequently, the issues may be more appealing to the unsophisticated but socially conscious individual. The experimental group, on the other hand, might have been somewhat perplexed by the complexity of these issues as well as satiated by the intensive treatment given them. The critical and obvious question, at this point, is whether or not we have dampened the experimental students' general interest in important public issues by presenting problems at too complex a level too early. The data in Table 11.23 would indicate not. While the control groups actually lost interest in those topics assessed by the general topics test, the experimental group gained significantly.

In summary, then, while the experimental group gained slightly on both the Project items and the general items, the control groups showed greater gains on the Project items but lost a significant degree of interest on the general items when compared to Project students. Assuming these differences to be more than chance findings, they might be interpreted to mean that adolescents between the ages of twelve and fourteen begin to gain in social consciousness, and when this occurs, begin to focus on social issues on which it is easy to polarize one's views. Since the experimental group gained only slightly on these issues, probably because of having dealt with these topics intensively, we might infer that such treatment modifies this interest. However, students of this age apparently tend to lose interest in more diffuse

TABLE 11.22

COMPARISON OF CONCORD EXPERIMENTAL GROUP WITH CONTROL SCHOOLS ON
 "PROJECT TOPICS" HEADLINES TEST WITH MEANS ADJUSTED
 FOR PRE-TEST SCORE AND INTELLIGENCE

School	N	Pre-test		I.Q.		Post-test		Adjusted Mean
		Mean	S.D.	Mean	S.D.	Mean	S.D.	
Concord	102	22.93	4.30	113.38	12.26	23.93	4.45	23.75
West	126	20.67	3.44	118.55	14.84	24.36	3.59	24.32
Wheat	187	22.64	3.71	116.66	10.83	26.73	4.06	26.51

Test of Differences Between Means

	Degrees of Freedom	F-Ratio	Probability Level
Total	2/410	19.56	.001
Concord v. West	1/223	4.62	< .05 > .01

and less salient issues. Here again the experimental group reversed this trend. While the measure is admittedly crude, the sample is relatively large, and we think this finding is well worth more intensive investigation to check out and refine our speculations.

Conclusions

The measures by which we attempted to evaluate the effects of the experimental curriculum are so varied that it would require a great deal of oversimplification to try and reach any general conclusions from the study. We might, however, make a few general observations at this point. This phase of the study has really been concerned with five questions:

1. Are general reasoning tests in the guise of "critical thinking" tests sensitive to the type of experimental curriculum provided the students in Concord?
2. Can students at the ages of 12, 13, and 14 learn to handle complex political and social issues at an abstract analytical level?
3. Assuming that students at this age can to some degree deal with political issues analytically, will this affect their behavior in an actual argument or persuasive situation?
4. If one gives priority to teaching students an approach to political and social problems, and only incidentally handles chronological history, does this seriously affect how much history the student will know at the end of the two year 7th -- 8th grade sequence?
5. How will the student's interest in political and social topics be affected by such an experimental curriculum?

Regarding question one, the answer is that general critical thinking tests are apparently not sensitive to the specific intellectual training provided the experimental group. The training the Concord students received was not in general reasoning skills, nor was it intended to be. We thought, however, that the training might have a more general effect on the student. This did not prove to be the case.

The answer to question two is apparently positive. Students were able to learn an abstract analytic system and apply it very well to relatively simple cases based on political and social issues. They have much more difficulty, however, applying the model to complex information. The answer to question three, however, is even more equivocal. In the Socratic Discussion Test, the gains of experimental students were statistically higher than those of

TABLE 11.23

COMPARISON OF CONCORD EXPERIMENTAL GROUP WITH CONTROL SCHOOLS ON
 "GENERAL TOPICS" HEADLINES TEST WITH MEANS ADJUSTED
 FOR PRE-TEST SCORE AND INTELLIGENCE

School	N	Pre-test		I.Q.		Post-test		Adjusted Mean
		Mean	S.D.	Mean	S.D.	Mean	S.D.	
Concord	102	25.75	4.10	113.38	12.26	26.66	4.08	26.50
West	126	23.29	3.57	118.55	14.84	21.17	3.58	21.63
Wheat	185	25.51	4.52	116.76	10.79	23.74	4.11	23.53

Test of Differences Between Means

	Degrees of Freedom	F-Ratio	Probability Level
Total	2/408	46.67	.001
Concord v. Wheat	1/282	41.65	.001

control students at only the .05 level. Moreover, the results of student-led discussions make this finding even more difficult to interpret. This is clearly an area which needs much additional research, especially in the development of a standard test situation from which one can make valid generalizations.

The results bearing on question four are most encouraging to those committed to an analytically oriented curriculum. Apparently the experimental group did not suffer any relative loss of historical knowledge because of the experimental curriculum's emphasis upon contemporary issues. Certain kinds of historical knowledge, moreover, were significantly better retained. It is important, however, to qualify this finding by pointing out the suspicious conditions within which the experimental curriculum operated.

The effect of the curriculum on the experimental student's interest in political and social issues was generally positive, although when compared to control groups he tends to show less growth in interest in those areas he has studied intensively and greater growth in areas he has not studied at all. This finding is difficult to interpret and requires more research to clarify.

In general we find the overall results quite encouraging. If it were possible to transplant the experimental curriculum into a school as a substitute for a conventional two year American history curriculum, the results certainly might warrant such a decision. The positive findings, however, should be carefully qualified. The teachers of the experimental curriculum were carefully selected because of their interest in developing such a curriculum, and all are generally very competent teachers. Approximately a third of the discussions were carried on in split groups of approximately 12 students. The students as well as the teachers were aware that they were part of an experimental program that was on trial. All of these factors theoretically give the experimental group an uncontrolled advantage. At this point it is probably safe to say that we need two very different kinds of research: basic applied research into the problems which have simply been raised by this study; and implementation research into the generality of this study when it is replicated in other schools under different conditions. We hope that the day is past when the educational engineer reads the literature of curriculum research looking for definitive and general answers to complex questions. It is probably more useful to look for interesting clues to educational practices with which he himself can experiment and make a contribution. It is highly doubtful that any findings coming out of a classroom laboratory will ever be applicable to all schools in all situations. The challenge is to demonstrate that students can learn certain kinds of intellectual operations or conceptual structures, and then to investigate the critical conditions which allow this to take place. This study should therefore be construed as but a precondition to the more important work of defining these critical conditions.

CHAPTER TWELVE

TEACHER STYLE AND THE ANALYSIS OF THE STUDENT-TEACHER DIALOGUE

Of the many problems that exist today in public education, one of the most serious is certainly the quality of classroom discourse.¹ Mayer concluded, for example, that if there was one universal criticism of education, it was that teachers talk too much.² Perhaps the problem is not so much the quantity of talk as it is the nature of the discourse in which the teacher engages the student, and the way the discourse is handled. Moreover, this problem is likely to be a persistent one. For although automation, audio-visual media, and large group instruction may succeed in supplanting the inefficient manner in which many teachers now accomplish the task of presenting and "teaching" information, we still face the problem of using the classroom dialogue for the development of more complex learning outcomes. This problem is especially salient for us, since we are concerned with teaching the student to apply a complex conceptual framework as he debates public issues, either with himself or others.

Educators in the last fifty years have tried to deal with this problem in a variety of ways. Dewey suggested that we concern ourselves with describing the nature of complex problem analysis, and apply our findings to the classroom. Other investigators have approached the problem empirically through systematic research into classroom behavior and teacher style. Unfortunately, one is forced to ask: Why has this considerable body of educational research yielded so little of direct practical use for raising the quality of classroom discourse? In this chapter we would like to deal with this question, and suggest a line of investigation which, we feel, may offer more fruitful results.

The Global Dichotomy

Much of the experimental work on teaching style has come out of the investigator's assumption that a particular kind of student-teacher dialogue is the most effective for teaching any content or intellectual process. Whether the learning outcomes are essentially attitudinal, intellectual, procedural, or manipulative, or a combination of these, it is assumed that there is a quality of style that runs through all good teaching. While much of this literature is general and not directly related to the teaching of political controversy, it has clear implications for the process of classroom discourse, which, as we pointed out above, is central to our thinking.

The most common quality presumably related to good teaching is usually presented in the form of a global dichotomy. It has been described in the

¹We wish to distinguish "teaching" from "discourse" here, although later we use the terms interchangeably. Teaching is the broader term, including the activities of lesson planning, reading of papers, and the like.

²Martin Mayer, The Schools (New York: Harper & Brothers, 1961).

literature under many names, some of which are listed below:

<u>Investigator</u>	<u>Dichotomy</u>
Anderson, H. H. Lippitt & White	Dominative-integrative Authoritarian-democratic- laissez-faire
Withall	Teacher-centered, learner- centered
Thelen & Withall	Teacher-centered, learner- centered
Perkins	Teacher-centered, learner- centered
Cogan	Preclusive-inclusive
Flanders	Indirect-direct influence

Anderson, who has reviewed the literature in this area, points out that:

In most teaching methods textbooks you will find reference to two basic teaching styles which can be called teacher-centered and learner-centered. (See, for example, Brandwein, Waston and Blackwood, 1958; Burnett, 1957; Burton, 1952; Cronbach, 1954; Grambs, Iverson, and Patterson, 1958.) Leadership is universally defined in terms of a hypothetical authoritarian-democratic dimension. Many labels have been applied to this dimension -- perhaps it should be called a dichotomy, for such it has been in practice: dominative-integrative, employer-centered -- employee-centered, teacher-centered -- learner-centered, therapist-centered -- client-centered, supervisory-participatory, directive -- non-directive; but the idea is basically the same.³

Anderson then presents an exhaustive critical review of the research bearing on the learning outcomes or productivity of groups or individuals in groups conducted in a "learner-centered" versus a "teacher-centered" way. His conclusion is enlightening:

To summarize the educational research reviewed in this article, eleven studies have reported greater learning for learner-centered groups, thirteen have shown no difference, and eight have found teacher-centered methods superior to learner-centered. It should be noted that while some investigations have reported a statistically significant difference favoring one method or other, it is doubtful if any of these differences are of practical or social significance.⁴

³Richard C. Anderson, "Learning in Discussions: A Resume of the Authoritarian-Democratic Studies," Harvard Educational Review, XXIX (Summer, 1959), p. 201.

⁴Ibid., p. 209.

If one speculates about the reasons for the inconclusiveness of the research coming out of the democratic-authoritarian or pupil-centered -- teacher-centered conceptualization of the classroom situation, at least four explanations present themselves:

1) The studies investigating this area lacked methodological rigor and adequate research design; 2) The teachers attempting to perform a particular teaching role did not adequately succeed; 3) The global conceptual dichotomy is not a sufficiently complex model from which to undertake fruitful research; and 4) Insufficient attention has been given to the problem of measuring learning outcomes at a higher cognitive level. Although Anderson claims that the studies he reviewed tended to lack methodological rigor, this is not true in all cases. Moreover, considering the number of experiments done, it is doubtful that one would not identify a significant trend in the research, regardless of experimental rigor, if the teaching variable under study were as important as its proponents claim. Regarding the success with which teachers or group leaders have been able to perform the "democratic" or "authoritarian" role, a number of investigators have dealt with the problem of describing operationally and quantitatively the difference between the pupil-centered and the teacher-centered style of teaching. The adequacy of the conceptual framework is another issue. The adequacy of a theory depends in part upon the quality of research which grows out of it. Does it lead to cumulative and useful findings? The student-centered school of thought has, in fact, produced much provocative research, including some of the most careful efforts to quantify important components of teacher behavior.

The emphasis of these studies is, however, upon a motivational approach to the analysis of teacher style. That is, observational instruments are designed to identify those qualities of teacher behavior which presumably create the social climate or conditions which will facilitate learning; they do not describe the cognitive or intellectual component of interaction. It is only later in the work of Smith⁵ that any serious investigator has attempted to analyze the cognitive process in which the pupil and teacher are mutually engaged and which presumably would lead to an intellectual model by which classroom discourse might also be evaluated.

This exclusive emphasis of so much classroom research on the motivational components of learning with little concern for the specific intellectual process being taught has, we feel, inhibited fruitful research in the teaching of explicit intellectual strategies for dealing with social studies content. It explains, moreover, why investigators in this area have tended to ignore the problem of developing more adequate learning outcome measures. If any intellectual process is facilitated by a particular motivational operation performed

5B. Othanel Smith and others, "A Study of the Logic of Teaching: The Logical Structure of Teaching and the Development of Critical Thinking, a Report on the First Phase of a Five Year Project" (Prepared at the University of Illinois, Urbana under Cooperative Research Project No. 258 (7257) of U.S. Office of Education, 1959 [Mimeographed]).

by the teacher, the problem of measurement is minimal in experimental work. Our own feeling is that the global motivational approach, in and of itself, has limited utility as an instructional theory. It is impossible to demonstrate this point, however, without presenting a number of specific studies which have, in fact, operated from this base.

Below are presented a number of specific studies which have operated from what we are calling a "motivational orientation." This is by no means an exhaustive review of the literature, which would simply duplicate the work of R. C. Anderson. Its objective is to point up a number of shortcomings which run through this theoretical approach.

While much less experimental work has been done on the analysis of the strictly intellectual or logical aspects of teaching, we shall consider a number of works in this area also. We shall then try to demonstrate how a more comprehensive view of the teaching act, involving motivational as well as logical characteristics of the teacher-pupil dialogue might serve as a useful framework within which to carry on empirical research on teacher style.

The Analysis of Teacher Style: the Motivational Approach

The research of H. H. Anderson. H. H. Anderson and his colleagues Joseph and Helen Brewer and Mary Francis Reed did some of the earliest work in the systematic analysis of classroom behavior.⁶ Anderson developed the theoretical construct domination-integration. Dominative behavior is defined as a mode of response to others, characterized by resistance to change, use of threat or force, or unwillingness to entertain or use the ideas of others. Integration is the opposite: willingness to change, use of voluntary control, willingness to use the ideas of others. Anderson's system consists of thirty-five major categories, each with sub-categories. There is a set of categories for describing teacher behavior and a set for student behavior. Listed below are some examples of Anderson's categories.

⁶H. H. Anderson, "The Measurement of Domination and of Socially Integrative Behavior in Teachers' Contacts with Children" Child Development, X (June, 1939), pp. 73-89; H. H. Anderson, "The Variability of Teachers' Behavior Toward Kindergarten Children," Studies of Teachers' Classroom Personalities, I: Dominative and Socially Integrative Behavior of Kindergarten Teachers by H. H. Anderson and H. M. Brewer (Stanford: Stanford University Press for the American Association for Applied Psychology, 1945), Applied Psychological Monograph, No. 6, Ch. II, pp. 43-108; H. H. Anderson, "A Revised Method for Recording Teachers' Dominative and Socially Integrative Behavior in the Schoolroom," Studies of Teachers' Classroom Personalities, II: Effects of Teacher Dominative and Integrative Contacts on Children's Behavior by H. H. Anderson and H. M. Brewer (Stanford: Stanford University Press, for the American Association for Applied Psychology, 1945), Applied Psychological Monograph, No. 8, Ch. I, pp. 15-32; H. H. Anderson, "Dominative and Socially Integrative Behavior at the Fourth and Sixth Grade Levels," Applied Psychological Monograph, No. 8, Ch. III, pp. 88-122.

Teacher's Behavior

- Domination, with evidence of conflict
 - Determines a detail of activity in conflict
 - Relocates child
 - Direct refusal or contradiction, evasion
 - Disapproval, blame, shame
 - Warnings, threats, reminders, promises, interrupts
 - Calls to attention
 - Punishment
- Domination with no evidence of conflict
 - Determines detail
 - Warning
 - Calls to attention
 - Makes gratuitous statement
 - method: statement
 - method: questions
- Domination in working together
 - Refusal, denial with explanation
 - Selects child
 - Approval of required work
 - Gives permission
- Integration, no evidence of working together
 - Extends invitation
 - Questions regarding child's interest
 - Statement regarding child's interest
- Integration with evidence of working together
 - Helps child define, redefine
 - Approval
 - Extends invitations
 - Questions regarding child's interest
 - Statement regarding child's interest
 - Admits responsibility

Child's Behavior

- Nervous habits (puts fingers in mouth, etc.)
- Looking up
- Leaves seat
- Undetermined child-child verbal contact
- Playing with foreign object
- Conforming
- Child domination
- Attacks status
- Response in recitation
- Answers spontaneously
- Answers when called on
- Seeks help
- Plans experiment
- Contributes to own problem
- Contributes to another's problem

Anderson's hypotheses are concerned primarily with the relationship of teacher-student patterns of interaction along his hypothesized dimension of domination-integration. Within this theoretical framework, Anderson has, over the years, accumulated a considerable body of findings. Flanders summarizes Anderson's findings as follows:

First, the dominative and integrative contacts of the teacher set a pattern of behavior that spreads throughout the classroom; the behavior of the teacher, more than any other individual, sets the climate of the classroom. The rule is that when either type of contact predominates, domination incites further domination, and integration stimulates further integration. It is the teacher's tendency that spreads among pupils and is continued even when the teacher is no longer in the room. Furthermore, the pattern a teacher develops in one year is likely to persist in his classroom the following year with different pupils. Second, when a teacher has a higher proportion of integrative contacts, pupils show more spontaneity and initiative, voluntary social contributions and acts of problem solving. Third, when a teacher has a higher proportion of dominative contacts, the pupils are more easily distracted from school work, and show greater compliance to as well as rejection of teacher domination.⁷

The very nature of Anderson's construct, domination-integration, requires a comprehensive category scheme in which task, procedure, and affective components of interaction are mixed under a single sovereign construct. Note for example that "disapproval," "determines detail," and "gives permission" represent affective, cognitive, and procedural aspects of interaction, and all are classified as dominative.

We question how useful it is to start with a global construct which then determines not only what components of the interaction process we choose for analysis, but also at what level of specificity we select categories, and how these categories might be organized. This may well place too great a burden on the construct itself, and obscure the complexity of the teaching act and a line of investigation which might lay bare this complexity. It is certainly possible that a teacher may be integrative in the area of classroom procedures, for example, but be quite dominative about the nature of knowledge, the validity of knowledge, and the nature of inquiry, i.e., in the cognitive domain. This kind of "contradictory" behavior is, we think, commonly observed in the classroom. Its effects, however, cannot be ascertained with the global dichotomy approach used by Anderson and his associates.

⁷Ned A. Flanders, Teacher Influence, Pupil Attitudes, and Achievement, (Prepared at the University of Minnesota, Minneapolis under Cooperative Research Project No. 397 of the U. S. Office of Education, 1960 [Mimeographed]), pp. 7-8.

Flanders' Work

Flanders undertook a study of how the spontaneous behavior of a teacher affects learning in the classroom.⁸ He developed an observational category system designed to pick up differences in teacher influence patterns in the classroom utilizing two theoretical constructs or the dimension of direct and indirect influence. "Direct influence by a teacher restricts the freedom of action of a student by setting restraints or focusing his attention on an idea. Indirect influence by a teacher increases the freedom of action of a student by reducing restraints or encouraging participation."⁹

Flanders' system used a classroom observer who classified statements into one of ten categories. Seven of the categories are used for teacher talk, two for student talk. One is used for silence or confusion. The ten categories are:

1. Clarify feeling constructively
2. Praise or encourage
3. Clarify, develop or make use of ideas suggested by students
4. Ask questions
5. Lecture
6. Give directions
7. Criticize or justify authority
8. Student talk -- initiated by the teacher
9. Student initiated talk
10. Silence or confusion.

Categories 1-4, according to the theory, characterize indirect influence; categories 5, 6, and 7 characterize direct influence. Entries in the system were made every three seconds. The observer also kept a record of change in type of classroom activity.

Three major hypotheses of the study were: (1) Direct teacher influence restricts learning when student perception of the goal is confused or ambiguous. (2) Direct teacher influence increases learning when a student's perception of the goal is clear and acceptable. (3) Indirect teacher influence increases learning when a student's perception of the goal is confused or ambiguous.

The measure used to get at indirect - direct influence is derived from a ratio of direct to indirect categories for each teacher (called an I/D ratio). Achievement is measured by objective test of factual information.

Differences in achievement between students taught by the high I/D ratio social studies and mathematics teachers were compared with the achievement of

⁸Ibid.

⁹Ibid., p. 12.

students of low I/D teachers. The results demonstrated that high I/D teachers were significantly more effective. Classes exposed to the most consistent indirect or direct patterns of influence, moreover, showed even greater differences.

While these findings are interesting -- studies in which teaching style has been objectively related to independent measures of learning outcomes are rare -- we find some difficulty with Flanders' theoretical interpretation. Let us look more closely at the construct from which his category system emanates, i.e., the operational definition of the construct, and the relationships of these operations to learning. Flanders explains the distinction between direct and indirect teaching as follows:

In essence, the difference between the direct and indirect patterns in this experiment is the central position of the teacher. The direct pattern is close to what the average instructor would call lecture with questions to check on the student's understanding of what was said. The indirect pattern shows a genuine effort to encourage student participation even though the strangeness of the teacher-pupil relationship and the highly organized program tended to inhibit the students.¹⁰

During a first year experiment, when members of the staff "practiced" the two styles, there were clear differences between indirect teachers and direct teachers in categories 1,2,3,4 (indirect categories), and 5, 6, 7 (direct categories). No statistical tests are reported. However, histograms are presented and the differences in frequency of categories 1,2,3,4 by indirect and 5,6,7 by direct teachers are striking. Inspection of this histogram shows that the direct teachers lectured twice as much as the indirect -- as a matter of fact, the direct teachers spent most of their time lecturing to the students, since from the theoretical point of view this was crucial. (The entire "discussion" period, however, lasted only fifteen minutes for each group.) In addition, direct teachers spent proportionately more time criticizing and justifying authority (categories 6 and 7). These theoretical differences were supported by the first year experimental results when the teacher practiced these styles of influence.

However, when we look at the second year experiment where teacher influence was not deliberately manipulated, i.e., field teachers were used, we find some interesting reversals in original findings. Indirect teachers spent more time lecturing. This difference, by an inspection of the data, appears to be consistent among all teachers. In addition, there are no consistent differences in the "asks questions," category or in the "student initiated talk" category (#9). The students talk proportionally less in indirect classrooms (#8 plus #9). There was consistently more confusion and silence (#10) in the direct classrooms. And if we look at the characteristic which

¹⁰Ibid., pp. 29-31.

differentiates the "indirect" teacher from the "direct" teacher, we see that the direct teacher "gives more directions," spends more time "justifying authority," and has more "silence or confusion" in the classroom.

In short, in the indirect classrooms, the teacher lectured more, while the students talked less; in the direct classrooms the teacher spent more time criticizing students, justifying authority, while there was more silence or confusion. We suggest that these findings can be easily explained without recourse to the abstract indirect-direct construct, which may in fact obscure the more interesting findings. The direct teachers he identified may have had nothing in common except that they spent a disproportionate amount of their time criticizing students and justifying their authority. This rather simple explanation, we think, accounts for many of Flanders' findings. The teacher, wrapped up in discipline problems, spends less time on the teaching task (hence lectures less). There is nothing in Flanders' theoretical position which would explain why the direct teachers should lecture less, and have more "silence and confusion." We do not have to explain away these findings. We suggest that Flanders has developed a scheme which will make gross differentiations among teachers who have discipline problems and those who do not, and this is not without value. He can, of course, call these teachers direct or indirect, but the label does not seem especially important or descriptive of what the indirect and direct teachers actually do.

We have tried out Flanders' system informally on practice teachers in three different classes. Only one apprentice, who was having "discipline problems" would have received many tallies in categories 7 and 8. The other teachers had approximately the same number in these columns. This was done very informally, yet the two teachers who scored approximately the same, appeared to be using extremely different teaching styles. Moreover, Flanders' categories did not pick up these differences.

In spite of the obvious merits of Flanders' work, we question the usefulness of the "direct-indirect" construct. Like Anderson, Flanders gives no consideration to the substance of the classroom interaction. It is a purely motivational scheme. Finally, we should note that Flanders' work bears on classroom learning defined only in its simplest terms; memorization and application of concepts in a multiple choice test. Our interest is, of course, in teaching the use of a somewhat more complicated conceptual framework. This is not a criticism of Flanders, but simply a point of fact limiting the relevance of his work to investigations, which focus on the teaching of complex thought patterns rather than simple recall of information.

The Hughes Study¹¹

The Hughes study is probably the most recent effort to capture through systematic observation the elements of dominative and integrative teaching.

¹¹Marie M. Hughes and others, A Research Report: Development of the Means for the Assessment of the Quality of Teaching in Elementary Schools (Salt Lake City: University of Utah, 1959).

Although the work appears to begin with a less dichotomized orientation than either Anderson's or Flanders', it settles down into the dominative-integrative framework offered by Anderson, and with categories which look much like Flanders'.

The "attempt to describe what goes on in American classrooms for children from the kindergarten through the sixth grade" is carried out in the following manner. Observers take extensive notes on actual classroom teaching. These notes are then compiled into specimen records which constitute the raw data to which a category is applied. The unit of analysis of the category system is called the "function." In determining this unit, the question is always asked, "What function did this verbal behavior and/or non-verbal behavior perform for a child, group, or class?" The specific behaviors of the teacher are then analyzed into thirty-one functions, which are subsumed under seven broad categories. These categories are, in brief:

1. Controlling Functions: Teacher activities designating the thing to which children are to give their attention, and the individual or group that is to do it; stating or asking for a recall of the proper and accepted way of doing things; judging who is right when there is a conflict of interest or dispute.
2. Imposition of the Teacher: Teacher projecting herself into a child's activity without being asked; doing routine clerical work for children (handing out paper, etc.); helping a child when he doesn't ask for it; giving information to the child which is not centrally relevant to the conversation, but which the teacher feels is "necessary."
3. Facilitating Functions: Neutral functions which facilitate the movement of content or activity underway. These activities include checking for information or clarifying a procedure.
4. Functions that Develop Content: This function involves such behaviors as repeating what the child has said, stating it in different words, and elaborating on what has been said; also asking the child for a personal opinion.
5. Functions of Personal Response: This function is a response to a personal problem expressed by the child. The problem may be asking for a piece of paper, asking for the teacher's attention, responding to a personal experience as it is recited by a child. Also included in this category is acknowledgement of a teacher mistake.
6. Functions of Positive Affectivity: This includes support, encouragement, solicitousness.
7. Functions of Negative Affectivity: This includes such subcategories as admonish, reprimand, accuse, threaten, ignore.

Having developed these categories, Hughes proceeded to demonstrate that elementary teachers in general use far too many controlling functions and impositions. Moreover, she found functions that develop content in short supply. Contrary to the author's intent this study does not seriously attempt to describe teaching in American elementary schools; rather it is meant to describe general criteria by which one can evaluate good and bad teaching. Categories four, five, and six are termed integrative responses, while categories one, two, and seven are considered dominative categories, after Anderson. At this point we are back to global dichotomies. Quoting Anderson in the report, Hughes presents two columns, side by side, labelled "Socially Integrative Behavior," subheaded "the creative growth circle"; and "Domination," subheaded "the vicious circle." The first two entries in each column are sufficient to indicate the tone and level of discourse:

Socially Integrative Behavior

The socially propitious human environment as it concerns human interacting is a mutually accepting circular phenomenon.

If one human being tends to make a propitious environment for another, the other tends to make a propitious environment for the one.

Domination

Domination is found in school-rooms, families, and factories. It has been a conspicuous part of international diplomacy of all time.

In the psychology of human behavior all uses of force, coercion, domination, shame, blame, guilt have one effect: the stifling of the creative process, the annihilation of originality. This is the environmental degrading of the quality of behavior.¹²

Reliance on this kind of thinking leads Hughes to a "model of good teaching" defined operationally and even quantitatively. The validity of the model is stated as the conditions necessary to produce the flexible inquiring student. The assumption is clear that integrative teaching will, in fact, produce a more useful, reflective, autonomous citizen.

Two problems are apparent in Hughes' work. First, contrary to some of Smith's assertions,¹³ it is doubtful whether the criterion of "good teaching" can be derived from the act of teaching itself. It is like deriving criteria for good skiing, running, or swimming by looking only at the form of the skier, the runner, or the swimmer. Presumably "good" form has a measurable

¹²Ibid., p. 202.

¹³B. Othanel Smith, "A Concept of Teaching," Language and Concepts in Education edited by B. Othanel Smith and Robert H. Ennis (Chicago: Rand McNally Co., 1961).

payoff in the speed or distance of the performer as the "form" of the teacher would have a payoff in the learning outcomes of the students. This assumption, however, remains to be demonstrated. Nor does Hughes explain the derivation of her quantitative definition of the "model" teacher. It is simply "developed from" the 129 classroom sessions from which specimens were taken. While Hughes' line of inquiry may serve the function of raising important questions about some rather obvious shortcomings and rigidities in the teaching role as normally performed, it offers little evidence from which to draw conclusions regarding the relationship between teaching style and learning outcomes. This is especially disappointing in view of the fact that the study was explicitly concerned with teaching higher cognitive functions. Second, it is disappointing that so little effort was made to define category 4, "Functions that Develop Content." One learns little more than that the chief characteristic of content development is the responsiveness of the teacher to the data the child places in the situation.¹⁴ The important question, from our point of view, focuses upon the logical nature of the teacher's responses. In our own work on the junior high school level, a noticeable intellectual characteristic of the seventh grader is his tendency to view the teaching situation as one of two kinds: the "rigid recitation" situation and the "opinion offering" situation. The difficulty is that in the "opinion offering" situation, the student does not expect his beliefs or opinions to come under serious question. He simply relates personal experiences about which the teacher may or may not express interest. The notion that the personal beliefs of the student should be subjected to critical evaluation appears to him somewhat unreasonable. Thus, while Hughes' model teacher may score very high on content development functions and positive affectivity, he may at the same time be reinforcing sloppy thinking, wild generalization, and egocentric conceptualizations and solutions to general social problems. Without more specific distinctions in the cognitive domain, one simply does not know.

The Analysis of Teacher Style: the Logical Approach

Studies of the cognitive processes involved in instructional discourse are rare in the experimental literature. The work of B. O. Smith and later that of Aschner are among the few studies of this kind.¹⁵

¹⁴ Marie M. Hughes and others, A Research Report, p. 69.

¹⁵ There are several studies which take the "logico-psychological" approach. The works of Shaver and Wright are somewhere between that of Flanders and Smith. Shaver presents categories which distinguish between two theoretical models of teacher style. He has only six broad logical categories; however, Wright attempts to set up both psychological and logical categories for the evaluation of mathematical instruction. James P. Shaver, "A Study of Teaching Style: The Investigation Through Systematic Observation of the Ability of Experimental Teachers to Conform to Two Models of Teaching" (Unpublished doctoral thesis, Harvard Graduate School of Education, 1961); E. Muriel J. Wright, "Development of an Instrument for Studying Verbal Behaviors in a Secondary School Mathematics Classroom," Journal of Experimental Education, XXVIII (Dec., 1959), pp. 103-121.

Smith questions a basic proposition which he says is widely held in pedagogical circles: namely, that teaching can adequately be explained and controlled on the basis of psychological knowledge alone. He holds that logic¹⁶ which establishes standards for discourse has largely been ignored. As empirical methods in psychology developed, the view that logic was supposed to deal with actual mental processes "came to be seen as gratuitous."¹⁷ It was apparent, Smith holds, that ordinary verbal behavior bore little resemblance to formal logic. Thus, it was not surprising that the young science of psychology reduced mental processes to sensations, images, associative mechanisms and the like.

Then, as experimental animal psychology replaced introspection, consideration of logical factors of intelligence were again swept aside along with the whole domain of cognition.... Thinking [in Thorndike, for example] was reduced to a series of stimulus-response actions occurring in accordance with the laws of readiness, exercise, and effect. And logic as the science of thinking was undercut.¹⁸

Although Dewey looked at teaching as guiding the learner through the process of discovery and verification according to his own reconstructed model of logical thinking, Kilpatrick and other educational psychologists who followed Dewey generally ignored the logical aspects of Dewey's thinking and emphasized the psychological or motivational needs of the learner. Today, we find no distinction between valid and invalid thinking for this distinction is unnecessary for a psychological analysis of the act of teaching. Smith argues that we must once again appeal to the norms of inductive and deductive logic if we are to advance in our understanding of the quality of classroom discourse.

Smith also questions another widely held proposition in educational theory -- that subjects of instruction do not differ widely in their logical demands. It is commonly held that the "problem solving method" is superior to all in developing critical thinking as well as other educational objectives. It is widely assumed, for example, that the ability to analyze an argument, examine experimental evidence, and analyze and criticize valuations can be taught and learned in geometry, English, history, or physics. Smith reports in a recent study that he was forced by the results of his own work to re-examine this assumption. Smith points to the possibility that differences among teachers may be due either to the different degrees of logical rigor by which they use higher mental processes or to differences in logical operations

¹⁶Smith does not use the term "logical" to mean formal logic. He uses it in the wider sense as the set of norms by which we judge the rigor of any inquiry: analytical, empirical, or ethical.

¹⁷B. Othanel Smith, "A Study of the Logic of Teaching," p. 1:2.

¹⁸Ibid., pp. 1:2-1:3.

built into the various subjects or to both of these. We should note that both Flanders and Hughes largely ignore the logical operations used by the teacher and the relevance of these operations to the content or problem under discussion.

Smith's research was directed at discovering the "logic of teaching" as exhibited in the classroom. More specifically, he and his associates tried to analyze classroom discourse into pedagogically significant units and classify them into logical categories. The long range hope is to instruct teachers to handle these logical structures and to measure the effects upon the thinking of students. After a fairly lengthy trial and error period which Smith describes in his report, he and his associates found that classroom discourse falls into two sorts of units; the episode and the monologue. The episode is a unit which represents a completed verbal transaction among two or more members of a class, while the monologue is a solo performance of a speaker addressing a group.¹⁹ Smith also found teachers used three different kinds of statements: "logically relevant tasks," "directive tasks," and "admonitory acts." He concerns himself with the logically relevant acts. Note that he ignores those acts which were of central concern to Flanders and Hughes. The logically relevant actions, which form the basis of his category system are as follows:

1. Defining
(four types)
2. Describing
3. Designating
(five types)
4. Stating
5. Reporting
6. Substituting
7. Valuating
8. Opining
9. Classifying
10. Comparing and Contrasting
11. Conditional inferring
12. Explaining
 - Mechanical
 - Causal
 - Sequent
 - Procedural
 - Teleological
 - Normative
13. Directing and Managing Classroom²⁰

¹⁹Ibid., ch. 3.

²⁰Ibid., p. 7:4.

After Smith had established some moderate degree of reliability for scoring classroom interaction with these categories, he attempted to determine what logical patterns if any could be found in teaching. Eighty-five transcripts of class sessions were scored. Smith then inspected the difference among operations in the various content fields and concluded that there are differences. He points out that teacher and subject are confounded, but it does seem likely that differences may exist in the extent to which the logical operations are employed from teacher to teacher and from area to area. "Adequate answers as to what this extent is, however, cannot be determined from our data, but await further studies designed more specifically to investigate these matters." Smith performs no statistical analysis for the differences in entries for teacher and subject, but relies on inspection for his generalizations.

In evaluating Smith's work, we should state that it is the first major effort, to our knowledge, to redress the balance between affective-procedural components in teaching style and the cognitive area. Smith's concern with the normative aspects of "critical" versus "uncritical" thinking points up the fact that we must not only find ways of encouraging the child to reflect and inquire; we must find ways of encouraging the child to reflect and inquire in a disciplined and intelligent way.

We should, however, point out what we consider as two weaknesses in Smith's pioneering work. First, in dismissing problems of control or procedure and problems of interpersonal affect (presumably to be investigated at a later date), we feel Smith does not see the full implications of his own effort. If we are to consider certain logical operations of the teacher as the same operations that it would be desirable for the student to perform, the cognitive description of teaching style also provides us with an evaluation instrument for measuring learning outcomes of students. And getting this far, it may well be feasible to investigate the complex relationships among the affective, procedural, and cognitive responses of the teacher and relate them to intellectual processes of the student, which can, in fact, be construed as learning outcomes. Second, Smith apparently begins with a general conception of "critical thinking" which is useful in analyzing a wide spectrum of curriculum areas. He is thus concerned with the extent to which teachers express these logical operations in the classroom, and the extent to which these operations differ across content. It seems obvious to us that in the analysis of teaching style, one might better begin with specific objectives the teacher has in mind, and proceed to investigate the extent to which the logical operations the teacher performs (or induces the students to perform) are appropriate for the objectives the teacher is trying to achieve. Thus, it may be more reasonable to begin with a species of problem and inquire into the type of logical operations required to clarify or resolve this type of problem, and limit the number and specificity of distinctions in the cognitive area to those which have some degree of demonstrable utility for this or that particular type of analysis. As the reader will recall from Chapter Ten, this, in fact, has been our own strategy in attacking the problem of measuring learning outcomes.

Aschner and Gallagher: A Comprehensive System

In her work with Smith, Aschner developed what she called the "neutral unit" or episode, which is the basis of unitization in Smith's category scheme. Using the same basis of unitization, Ascher, in conjunction with Gallagher, has developed a more comprehensive instrument by which classroom behavior can be analyzed and quantified. The two criticisms of Smith's work suggested above -- the lack of concern with procedural and affective responses, and lack of focus on a specific problem area -- are less applicable to Aschner's work. The scheme was worked out to deal with a specific problem: To identify the kinds of thinking carried on by gifted children. It is more comprehensive than Smith's, although affective-procedural responses are conceptually handled under a single broad category called "routine." Routine responses are, however, broken into a number of more specific and meaningful subcategories.

Aschner and Gallagher's work grows directly out of Guilford's conceptualization of cognition. While this is certainly appropriate in the light of their objectives, one has trouble translating Guilford's categories into educationally significant terms. Strangely enough, Aschner talks about cognition in very much the same way as Hughes. The teacher who performs "content development" in Hughes' terms is the same teacher who elicits "divergent thinking" in Aschner's terms.

While we are impressed by the comprehensiveness of Aschner's work, as well as the value of grounding it theoretically, we are a little at a loss to know how to evaluate the quality of thinking performed by the students (or the teachers) as that thinking is defined by these categories. Certainly divergent thinking is not a good in and of itself. For example, there may well be a phasing out of the four aspects of cognition upon which the scheme is based:

the student must have a fund of information before he can deal with an issue (memory);

the student considers various alternative responses to the issue (divergent thinking);

the student evaluates these responses according to their logical soundness (evaluative thinking);

the student develops a specific solution (convergent thinking) taking into account many of the other alternatives (divergent thinking).

The point is that the quality of a logical operation must be determined partly by its relationship to the structure of a total argument. And the structure of an argument may well vary, depending upon the domain of the world with which it is concerned (art, politics, physics), and the way one looks at this or that particular kind of problem (as an aesthetic problem, a scientific problem, an interpersonal problem, etc.). We want to be clear in stating that our criticism of Aschner's work rests on the assumption that it was designed to have instructional implications. No explicit attempt has been made to relate her observational tool to curriculum decisions, although this is

CATEGORIES FOR CLASSIFYING THOUGHT PROCESSES IN CLASSROOM VERBAL PERFORMANCE²¹

I. ROUTINE

Management
question
procedure
directing
aside
nose counting
feedback
seatwork
footwork

Structuring
self
other
class
pedagogic

Verdict
favorable
unfavorable
pending

Agreement
affirmative
negative

Monitoring

Self-Reference
dunno
self report

Muddled

Humor

Scribe

II. COGNITIVE - MEMORY

Recapitulation
quoting
repetition
recounting
review
explaining

Clarification
meaning
qualifying
correcting
adding

Factual
fact stating
fact detailing

Pedagogy

III. CONVERGENT THINKING

Translation

Association Convergent
comparing
linking
general statement

Explanation
narrative
rational
valuation

Conclusion
summary
logical
pedagogical

Convergent Problem
answer
answers
solution

IV. EVALUATIVE THINKING

Rating

Viewpinion
probabling
choosing
qualifying

Evaluation
either-or
choosing
qualifying
judging
counter-judging
pedagogical

V. DIVERGENT THINKING

Elaboration
structured
free

Association Divergent
structured
free

Implications Divergent
structured
free

Divergent Problem
divergent implications
divergent solution

²¹M. J. McCue Aschner, "The Productive Thinking of Gifted Children in the Classroom," Paper delivered to the American Education Research Association, Chicago, February 24, 1961 (Mimeographed).

clearly the long range goal. We feel that "productive thinking" may well have a different meaning, depending upon the nature of the problem one is trying to handle and the context in which it is handled. From this point of view, that aspect of teaching style concerned with the logical operations performed by the teacher is inextricably tied into a series of normative assumptions about the nature of problems and the nature of valid and appropriate solutions.

Some Concluding Remarks about Research in Teaching Style

In looking back over the studies in teaching style reviewed above, we wish to emphasize three points: (1) The analysis of teaching style has tended to follow from motivationally oriented theories of teaching, which have stressed the procedural and affective components of teaching to the neglect of the cognitive area. To the extent that the cognitive area has been considered, a very simple dichotomy has been suggested along the lines of "teacher broadens and facilitates cognitive activity by student" versus "teacher constricts and inhibits cognitive activity by student." (2) In the content analysis system of B. Othanel Smith in which cognition is central, there is insufficient stress placed on the fact that the analysis of good teaching is precisely the same kind of analysis required to evaluate the verbal behavior of the student. Instead of stressing this fact, Smith goes to great lengths to demonstrate that teaching can be analyzed independently of learning outcomes. (3) The empirical work done on the description of teacher style proceeds from the general context of "classroom interaction." Clearly classroom "discussion" means many different things to teachers in a given subject matter area; and this meaning can be multiplied by the number of different subjects taught in the schools. It can mean drill on memorized facts; application of concepts; the settling of procedural problems in the classroom; analysis of complex societal, mathematical, or scientific problems; or the defense of a position with respect to a problem. Different classroom discussion situations, then, have different objectives, which may determine how they are set up and organized by the teacher. Before one can evaluate classroom interaction, in general, one would first do well to consider the instructional objectives of the person who set up the situation.²² Our objectives have, of course, been described in some detail in Parts One and Two; and it is within the framework of these objectives that we have conceptualized a way of talking about teacher style.

²²It should be clear, of course, that before we can talk about classroom discussion at all, we must first distinguish between "mass media" and the face-to-face dialogue. Mass media include TV, film, lectures, or any instructional situation in which there is a one-way flow of information. The dialogue says exactly what the term means.

Our own work has concentrated upon the analysis of patterns of cognitive discourse which we think are useful in the classification and resolution of political problems. From our point of view, this is the major task of the dialogue.

A Multidimensional Analysis of the Act of Teaching

In our attempt to conceptualize a more adequate theoretical framework from which to look at the teacher-student dialogue, we started with the distinction first set forth by Bales' Interaction Process Analysis²³ between social-emotional and task responses. We then made a further breakdown in the task area between those responses which attempt to structure the task in the immediate interpersonal situation, and those responses which deal with the political problem under discussion. We called the former procedural responses and the latter cognitive task responses. These distinctions have borne up well under systematic investigation, as suggested by Anderson.

There are probably many variables of leadership and group life. Hemphill (1949), after analysis of some five hundred existing groups, arrived at sixteen descriptive variables. Factor analysis of Hemphill's variables (Gekoski, 1952), as well as factor analyses by Carter and Couch (1953), yield three dimensions which the present author likes to interpret as follows:

1. The Affective area. Includes degree of interpersonal warmth or coolness, tension or relaxation; degree of antagonism or solidarity. Opposes friendly pleasant exchange to hostility and constraint.
2. The Procedural area. Includes statement and control of agenda, control of communication, division of labor. Refers to amount of structure and degree of organization as opposed to degree of disorganization or looseness of structure.
3. The Task area. Includes quantity and quality of work done, units produced, ideas presented, solutions considered. Opposes solution reached, problem solved, high performance maintained to no solution, problem unsolved, poor performance. Productivity versus sterility.²⁴

We then analyzed the affective-procedural-task trichotomy into a larger number of more specific dimensions. These were selected on the basis of two criteria. First, they seemed logically related to the political controversy discussion

²³Robert F. Bales, Interaction Process Analysis (Cambridge, Mass.: Addison-Wesley Press, 1951).

²⁴Richard C. Anderson, "Learning in Discussions," p. 210.

model with which we were centrally concerned. And second, certain components of the discussion process have been used in other studies and been found to bear a relationship to student learning. Below is a brief statement of these dimensions of classroom discourse. We feel that they are, at best, minimally adequate as a conceptual model to handle the complex phenomena that occur in the teaching dialogue.

A Set of Dimensions for Describing the Act of Teaching

A. Cognitive or task dimensions.

(1) Statement posture.²⁵ The term "statement posture" refers to whether the teacher or student is making a positive assertion, asking a question, questioning or challenging a statement made by another individual, or expressing self-doubt, either by questioning one's own statement or one's competence to make a statement. Statement posture was a major discriminating factor in Flander's work, and has obvious implications for describing important distinctions in teacher style.

(2) Discussion posture. Individual statements generally reflect the speaker's attitude toward or way of handling information.

We have been concerned with two aspects of the teacher's approach to the content of a discussion. If he discusses the content as problematical with no definite "right" answers, we would term the discussion dialectical. If the teacher discusses the content, through either statements or questions, with the assumption that he is dealing with definitive truth, that is, certain knowledge which is either present or available, we would term the dialogue (or, monologue, if in lecture form) as descriptive. Stated another way, while both descriptive and dialectical teaching may assume a problem or question before the class, the dialectical posture necessarily does. In the latter case, it is also assumed that the student has some degree of autonomy in the way he deals with the question or problem. In the pure sense of the dialectic, it is assumed that the outcome of discussion will not be the sum of the teacher's contribution, but a result of teacher-student interaction producing a product which cannot be wholly accounted for as coming from one or the other, or from a summing of the contributions of both.²⁶

A second dimension of discussion posture, which may be applied in either a dialectical or descriptive situation, is the extent to which the discussion or problem under consideration is centered around the substantive issue

²⁵The reader should understand that the terms used to denote these dimensions are arbitrary stipulative definitions.

²⁶We would emphasize again that this analysis is meant to handle only the general domain of political controversy; it has not been developed to handle the analysis of teaching in subjects such as mathematics, science, and literature. But it might well be appropriate in these areas as well.

described by a commonly agreed upon set of categories, as opposed to an analytic approach to the problem, in which an effort is made to explore new ways of looking at or structuring the nature and scope of the problem so that its substance can be more adequately handled. For example, a group might discuss whether the United States should support or withdraw from the United Nations. The argument might focus on whether or not UN actions are a beneficial extension of the United States foreign policy of collective security. At this point collective security might constitute the substantive issue under discussion. One member of the group might then suggest that the justification of the UN could be discussed in several contexts: (a) as a primitive instrument of world law which superceded the concept of "foreign policy"; (b) as a propaganda safety valve which allowed various governments to gauge world opinion of the actions of individual states; (c) as a tool of foreign policy for individual nations. Although a statement such as this might broaden the context of discussion and suggest a more complex framework within which to deal with the issue, it makes no judgment itself concerning whether or not the UN is a worthwhile institution. We call such statements analytic rather than substantive, since they do not focus on a suggested solution to the point under controversy, but rather on the conceptual context within which the point under controversy is to be handled.

In summary, then, we are suggesting two dimensions of discussion posture:

descriptive vs. dialectical;
substantive vs. analytic.

(3) Statement types. Another level of cognitive analysis concerns the specific type of statements used in classroom discussion. Does the teacher, for example, give or ask for a high frequency of value judgments; or is he basically concerned with either giving or asking for factual claims about reality? Does he spend a great deal of time repeating and summarizing what has gone before in the discussion? If he requests value judgments from his students, are these requests directed at the level of general or specific value judgments? In our systematic analysis of classroom discussion we have, for example, differentiated the following statement types: general values, specific values, general legal claims, specific legal claims, factual generalizations, specific factual statements, sources of evidence, statements of clarification, statements of repetition, definitions, and analogies or "cases."

(4) Logical or intellectual operations. A fourth level of analysis deals with the logical or intellectual operations performed by the teacher or student as each attempts to break down or develop a policy decision. In general terms, some of these operations are: problem differentiation and focus (is the problem mainly ethical, factual, or definitional?), relevance testing (is a statement relevant to the discussion and to the particular problem presently being considered?), dialectical strategy (the use of specific instances to contradict more general statements, and the use of qualifying statements to cover these inconsistencies). Also the complexity of a cognitive episode should be kept in mind. An

argument or "brief" might consist of two vaguely related statements (e.g., the United States should continue nuclear testing because you can't trust the Russians), or it might consist of a complex act of explicitly inter-related propositions.

Defining teacher style in terms of statement types and intellectual operations is a relatively new idea, and is undoubtedly of critical importance in the area of political controversy analysis, where both the efficiency of the teaching situation and learning outcomes of students may be defined with the same intellectual or logical concepts. The extent to which the teacher tends to be consistently concerned with the relevance of statements made in discussions; the extent to which he gives and demands evidence for claims; the extent to which the teacher chooses to treat analytic problems as arbitrary matters of convention by a rather free and easy use of stipulative definitions, or feels constrained to handle definitional problems within the network of existing meaning of words: These kinds of decisions on the part of the teacher may well be reflected in the way the student comes to handle similar problems, and thus may also be considered as evidence of student learning.

B. Social-emotional dimensions.

(1) Interpersonal affect. Another aspect of classroom behavior which must be taken into account in an adequate description of teacher behavior is interpersonal affect, i.e., the positive or negative feelings communicated in the interaction process. The dimension is so obvious and has been discussed frequently enough to require little explanation. We should point out, however, that it is possible to discriminate at least two different types of affective expression:

support vs. antagonism; and,
tension vs. tension release.

The behavioral cues for the first dimension require little explanation. The most obvious example of behavior indicating tension release is laughter. Evidence of tension is, however, harder to infer from observable classroom behavior. It can, however, be operationally defined by such actions as pencil tapping, pacing the floor, and nervous laughter. Of course, it is assumed that tension underlies, and is the cause of, the more boisterous laughter which is taken to be indicative of tension release.

C. Procedural dimensions.

Procedure refers to direct efforts to control the immediate interaction situation, e.g., giving directions for the conduct of a discussion, giving assignments, handling discipline problems. In our own work, the only procedural distinction we have drawn is between task-oriented procedural acts and those mainly concerned with deviance control. For example, telling the student to take out his book would be task-oriented;

telling the student to stop talking to his neighbor would be deviance control.

Even with the simple distinctions suggested here, the systematic analysis of the teaching act becomes a very complex business. Operationally, we can ask the following questions as we observe each statement in a discussion.

1. Is the act (or interact) a statement, a question, the questioning of another statement, or an expression of self doubt?
2. Is the act cognitive or procedural? If it is cognitive, we can ask a number of more specific questions:
 - a. Is the act descriptive or dialectical?
 - b. Is the act substantive or analytic with respect to the problem under consideration?
 - c. What type of statements are being made: value judgments; factual claims; definitional statements; clarifications, etc.?
 - d. What intellectual operations are in evidence: discriminating among statements with different degrees of relevance; generalizing from specific cases; specifying the meaning of a general statement; qualifying a general statement to take into account inconsistency; etc.?
3. If the statement is procedural, is it simply to facilitate the completion of the task, or is it an attempt to control deviant behavior on the part of group members?
4. Are there observable affective overtones to the cognitive or procedural message? If so, are these overtones antagonistic or supportive? Is tension or tension release present?

A Broader View of Discussion Style

Although we can dissect the dialogue into any number of appropriate dimensions, and identify teaching behavior along these dimensions, we recognize, at least intuitively, that there may well be certain general styles of teaching which are commonly observed. In dealing with the problem of describing these general styles, an important question arises: Which of the various dimensions set forth should be considered central, and which should be considered secondary in the definition of such styles? In determining the effect of a general style of teaching on learning outcomes, for example, is it more important that teachers be alike in the posture used, in the intellectual operations performed, in the amount and direction of affectivity shown, or in all three areas? Experimentally, there are at least two strategies one might use in answering this question. One might obtain a large sample of teachers, give them the same materials to teach, ask them to direct the class's attention

toward the analysis and clarification of controversial political issues, observe the range and profiles of behavior that follow and relate these to learning outcomes. Or, one might define two or more particular styles of teaching along those dimensions of behavior that one felt were critical, train teachers to play the styles, observe to what extent the dimensions of teaching not included in the definition were also affected by the different styles, and again determine the relationship of behavioral dimensions to student learning. In the long run, we are saying that our ability to identify and differentiate teaching styles is justified by the kinds of predictions we can make about learning outcomes. No matter how different two teaching styles may appear to an observer, with or without the tools of systematic observation, they are "significantly different" from an educational standpoint only if they lead to different learning outcomes.

In our own work we have chosen the latter course: To define styles according to certain conceptual dimensions, train teachers to play these styles, and observe what differences occur, if any, in behavior as well as differences in the learning outcomes of students. We might illustrate this approach by describing two style models we have investigated experimentally. We have called these styles "recitation" and "Socratic" teaching.

Recitation teaching. Probably the basic characteristic of recitation teaching is the teacher's attitude toward and control of knowledge. He provides through reading assignments -- usually a text -- and through his role in class discussions the correct information which the students are to know. The teacher expects the student to respond when called upon to fill in the sequence of information which the teacher wishes to develop in class. This involves mainly relating personal experiences or repeating or paraphrasing what the teacher or a text has said, although it may require the use of some independent thought, such as reorganizing previously read material or applying it to a new situation. In essence, then, evidence of learning is contingent upon the extent to which the student can respond to questions regarding information given him by the teacher, either in class or through texts and other media.

Socratic teaching. Socratic teaching, as we have conceived it, is clearly adversarial. When the center of discussion is a controversial political topic -- as in our work -- Socratic teaching requires that the student do more than describe the controversy in the terms in which the teacher, or assigned materials, have presented it. Rather, the Socratic teacher requires the student to take a position on the issue, state that position, and defend it. Here the emphasis is not only on knowledge provided by the teacher, as background for the discussion, but on the process by which the student arrives finally at a decision about the topic under consideration, on the careful consideration of alternative decisions, and on the utilization of analytic concepts and strategies, regardless of the position which is finally reached.

Describing General Styles Dimensionally

While these brief descriptions of the two styles suggest that we "know" what are, in fact, the differences between Socratic and recitation teaching, when one begins to perform as a teacher, the general descriptions very quickly prove themselves inadequate. It is this fact that has led us to the more microscopic dimensional analysis. We found we needed a more precise and reliable way of describing the teaching performance than casual recall based on anecdotal records. It should be clear, however, that the microscopic analysis does not replace the general style description. Theoretically, it will simply provide us with a more precise way of defining "natural" teaching styles.

To illustrate how these two levels of description are related, below we have defined some of the similarities and differences between Socratic and recitation teaching according to the dimensional analysis given previously.

Statement posture. Recitation teaching is characterized by a high frequency of "stating" and "question asking," with less emphasis on "questioning" or "expressions of self-doubt." With its adversarial nature, Socratic teaching shows a much higher frequency in "questioning" responses.

Discussion posture. Recitation teaching tends to be descriptive; it is assumed that the truth of the situation is available and that one has only to present and clarify information or an analytic structure by which information can be organized. The attempt to push the student toward a personal decision in which values are at stake is inappropriate for this style. The Socratic style is clearly dialectical. It assumes that the problem can be clarified only in an adversarial context in which various points of view can be presented and defended.

Both recitation and Socratic teaching may emphasize either substantive or analytic responses. The recitation teacher, however, will be concerned with the substance of the issue only insofar as he is interested in clarifying it and presenting the correct position. In the process of clarifying, he may well be analytic, asking that his students explain, "What are the possible ways one might look at this problem?" The concepts which the students apply to this analysis will be those given them by the teacher as correct. The recitation teacher will often avoid giving an answer to the substantive issue, feeling that here knowledge is not certain so he should only attempt to present his students with the knowledge and concepts which they should use in resolving the problem outside of the school. In using the Socratic style, on the other hand, the teacher focuses directly on the substantive issue and possible answers to it. From time to time, however, the Socratic teacher may depart from the dialectical posture and treat the problem analytically in order to facilitate conceptualization of the problem. The structure used for such an analysis by the Socratic teacher may have been taught in a descriptive manner, or if the teacher maintains a "pure" style throughout,

will likely have evolved from a dialectical discussion of the proper framework for viewing controversial political issues.

Statement types. Recitation teaching involves mainly factual claims. The teacher tends to deal with descriptions of reality, rather than with the truth or goodness of these descriptions. The recitation teacher in maintaining the sequence of his lesson -- to the extent that it is successfully programmed -- tends to proceed by gradual, well-related steps, using a great many summarizing, repeating, and focusing statements. The socratic teacher, on the other hand, in dealing with political controversy, tends to ask for value judgments and to challenge them. Factual claims will deal mainly with the background to the issue and as support for value statements. There is also a sequential factor involved in socratic teaching. It is likely to involve an emphasis on factual statements and questions as the discussion gets under way with the emphasis on value issues coming later. Recitation teaching, as far as we have been able to ascertain, does not reflect any sequential pattern.

Intellectual operations. Except for the fact that the socratic teacher is engaged in an adversarial role, there is no reason to expect that socratic teaching will evidence any different patterns of critical thought than will recitation teaching. The recitation teacher may question relevance, ask for or give evidence, or ask his students to qualify statements. However, our experience is that the socratic teacher does use more of these operations dealing with political controversy. It is also likely that because the teacher and students are engaging in controversy rather than talking about it, the intellectual episodes are more complex than in recitation teaching.

The procedural dimension. Procedurally, we have considered the teacher as the channelizer of a dialogue between him and the students, regardless of which of the two styles he is using. This, perhaps, points up our cognitive bias. It seems to us, however, that it is more fruitful to investigate procedural style differences within a limited context of general classroom procedures. If, for example, the teacher ordinarily withdraws to the back of the room to let students run the discussion, this is less a matter of style than general pedagogical methodology. The concept of style for us assumes a common curriculum, common objectives, and a common procedural context. Variations that occur with this much held constant, then, we would denote "style." The number of times a teacher is required to act procedurally to control deviance, for example, would be a matter of style.

The affective dimension. Theoretically, one may not link high or low affectivity with any particular style. In practice, the socratic discussion tends to be highly charged with negative affect because of the open controversy on the cognitive level, which spills over into the affective domain.

Utility of a Dimensional Analysis of Teacher or Discussion Style

We are suggesting here, as indicated by our previous discussion, that a dimensional analysis of teacher style might be useful in a number of ways. First, it allows us to define general teacher style more precisely. Instead of beginning with polar constructs such as student-centered -- teacher-centered or democratic-autocratic, for example, and developing categories to differentiate the poles, we define a number of dimensions of teaching, and then proceed to identify the characteristics of a teacher whom one would judge to be "student-centered," "democratic," "truth-seeking," or "Socratic." Thus, we are not trapped with a conceptualization of teacher behavior that works with only limited types of teaching.

Second, it allows us to describe variations in teaching styles which may or may not significantly affect learning outcomes. Instead of setting up styles in terms of single dichotomies and defining these operationally with "good" and "bad" categories, we can attempt to identify commonly observed styles of teaching, and determine which elements in the style, when varied, will or will not affect learning outcomes. Thus we can develop an empirical approach to the evaluation of teacher style. In the Socratic style, for example, we can ask whether the affective overflow which occurs when the student feels a personal threat to his ideas is a necessary part of the style, or is a low affect Socratic just as effective? In other words, is it the logical performance of the teacher which makes the learning more or less effective (if it is), or is it the logical performance plus the affective charge injected into the discourse.

From a research point of view this methodological approach allows us to investigate a number of interesting problems:

1. Is it possible for individual teachers to manipulate their behavior and play more than one style, or must we identify teachers who have natural styles we wish to investigate and compare?
2. If it is possible for teachers to manipulate their style, and play more than one style, what are the variations that occur within a single style, and are there overlapping areas in which one style cannot be differentiated from another?
3. Assuming that students were subjected to two quite different teaching styles on a systematic basis over a sustained period of time, under what conditions would this difference result in any different learning outcomes?
4. What is the relative importance of the intellectual or other personal characteristics of students, the style of teaching, and the personality of the individual teacher in effecting learning outcomes?

5. Is it possible to develop an observational instrument to describe not only the general style factors of the teacher, but also sufficiently complex intellectual processes that occur in the classroom, so that it can be used to assess learning outcomes?
6. How difficult is it to train teachers to operate consistently from a particular teaching style, and are there personality factors which make certain teachers incapable of learning some styles of teaching?

In our own work we have focused mainly on the first five questions, and have simply sidestepped the last question by selecting teachers for our research who were sufficiently flexible to teach in a variety of ways. It should be noted, however, that most of the people who have taught for our experimental program could handle socratic teaching with ease, and had more difficulty avoiding challenging the student's value position and initiating an adversarial dialogue. The problem of teacher training, therefore, would seem to be not only whether or not the teacher is sufficiently quickwitted and flexible to perform in a dialectical context, (the major problem of beginning teachers), but also whether or not he can tolerate giving doses of "truth" when they are called for. And the problem here may stem from basic temperamental differences among people, something professional educational programs can probably do little about. (Analogously, there may be temperamental as well as intellectual characteristics which cause some physicians to specialize in surgery and others to specialize in internal medicine.)

The Two Styles Investigated

In the course of our work we have focused mainly on the analysis, conceptual differentiation, and quantification of two complex teacher styles. We have called them the Recitation-Analytic Style (RA) and the Socratic-Analytic Style (SA). Since we have already described the distinctions between recitation and socratic teaching, the following brief operational definitions of the two styles should be adequate.

Recitation-Analysis

Recitation-Analysis (RA) as we define it involves the following steps.

1. A controversial case is read aloud.
2. The students are first asked to describe what took place in the case in their own words. This step is called descriptive recitation.
3. Students are then asked to tell why the situation in the case might be considered either good or bad. Students are not to give a personal evaluative reaction. They are to begin the analysis of the case by using the thought process concepts. This involves:

- (a) Describing the conflict in the case according to general social and political values;
 - (b) Telling what important factual assumptions relate to the conflict;
 - (c) Telling what analogies one might give to put the values in this particular case in perspective;
 - (d) Telling what consequences one might anticipate if any of the possible decisions suggested in the case were made;
 - (e) Telling whether any words in the cases might create special definitional problems;
 - (f) Telling how one goes about dealing with these definitional problems;
 - (g) If there are questionable factual assumptions in the case, telling what evidence the students might gather to support a particular claim, and what evidence already in the case supports any particular claim; and,
 - (h) Telling whether there are any rhetorical devices, such as the use of loaded words, which might influence one's judgment.
4. Students then summarize what particular alternative decisions are available in this particular case, and what important considerations one might take into account in arriving at an intelligent qualified decision.

In the initial stages of RA teaching much of the analysis and application is simply explained to the students by the teacher. As the treatment proceeds, however, the students carry out more of the analysis without the teacher's help, and they are simply told whether or not the analytic concepts are correctly applied.

Socratic-Analysis

Socratic-Analysis (SA) involves the following steps.

- 1. A controversial case is read aloud to the students. (Same as RA).
- 2. The teacher first asks students to recite events in the case to assure himself that the major points are understood.
- 3. The students are then asked what their personal feelings are about the situation described.
- 4. After several students have given an evaluative reaction, and a description of their position justifying the reaction, the teacher

begins a period of intensive questioning, concentrating on one student at a time.

- (a) If a student simply says, "This situation is bad," the teacher then asks him to explain what is so bad about it. The student then usually repeats his position.
 - (b) The teacher appears enlightened and then verbalizes for the student, if he has not done so for himself, a general social value which supports his position, or the violation of which disturbs him. The teacher then attempts to persuade the whole class to agree with the student's position. Usually most students will agree, since controversial cases tend to be loaded in one direction.
 - (c) The teacher then presents a second situation (telling it as an informal story) which illustrates the same values in conflict as those described in the case, except that the value loadings are reversed. For example, if a Negro sit-in case is presented which stresses equal opportunity as good and ignores the value of property rights, an analogy might be used to present the confiscation of American property in Cuba by Castro, which would make the importance of property rights more salient.
4. Usually students are unable to resolve the inconsistency between their feelings about the case and the analogy. They know that both involve the same value conflict, but they are sympathetic to equality in one and property in another. When the students have reached some degree of agitation, the teacher suspends the intensive questioning and asks the student to deal with the contradiction analytically. He shows the student how to qualify his position, how to question the factual assumptions the teacher commonly makes, and how to deal with the definitional licenses the teacher takes. This step is much the same as step three in RA teaching.

SA teaching is an attempt to teach the student to think on two levels simultaneously while he is discussing a case. The student must deal with the problem of persuading the teacher that he is taking a reasonable position with respect to the problem presented in the case. This is the argumentative level. At the same time, he must see that this can be done more effectively by using the political values and critical thinking concepts taught earlier. This is the analytic level. Initially, even after the students are taught the analytic concepts, few see their relevance to the controversial discussion. They tend to argue "blindly" in the sense that they have no conscious framework by which to analyze the nature of the problem. The teacher can argue more forcefully, because he can anticipate how to maneuver the student into unavoidable inconsistencies. We attempted to teach this two level consciousness by bringing the argument to a halt after a period of frustration, and asking the student to apply the concepts to this particular argument. If the student is unable to do this, the teacher "explains" the argument in terms of the analytic

concepts. ("I am attacking your value by the use of this analogy. You can counter this analogy with one of your own, qualify your position and make the analogy less relevant, or you can find some important difference between the situation in the analogy and the situation in the case, which also makes the analogy less relevant.")

After some analysis the argument may or may not resume. This depends on the timing and the enthusiasm of the class. While initially, the class is a patchwork of two phases: Socratic questioning → analysis and application of concepts → Socratic questioning, the teacher works in the direction of less and less analysis as the students become more proficient in using the concepts to guide their arguments. The tone of the class theoretically should move from a strictly adversarial one to one of a mutual search between teacher and students for an intelligent position in the case.

In comparing the socratic element of SA teaching with RA teaching, there are two striking differences. First, socratic teaching demands relatively long interchanges between student and teacher. The student cannot be led into an evaluative or definitional inconsistency in one or two statements. The teacher must first establish what the student's position is, suggest exceptions through analogies or counter cases or through contradictory evidence, and counter the student's defense. RA teaching, on the other hand, deals with the problem presented in the case only from an abstract position, never from the point of view of the student's personal commitment. The teacher would ask, for example, "What are some of the ways we might deal with the problem presented in this case?" Several different students can then summarize alternative positions in a sentence or two. The teacher might then ask, "What is the relative importance of the two conflicting values in this case when we consider this particular position....?" Although one position may be given sustained analysis in RA teaching, no person is asked to defend the position, and thus, no person becomes involved in a protracted dialogue with the teacher.

Second, as we have already noted, the level of affect in socratic teaching tends to be high. In RA teaching the student is asked only to apply a set of analytic concepts to a particular case. The student has no commitment to the analytic content other than its existence as something that the teacher has taught him. Nor is he asked to state a commitment to a position in the case. Misapplication of a concept requires an immediate corrective response from the teacher. In a socratic discussion, however, the teacher has no immediate responsibility for the inconsistencies or "holes" in the student's position in the argument. The student has autonomy to accept any position he chooses as long as he can withstand the teacher's probing. The student becomes excited and often agitated because he sees that two values, which he himself holds and is committed to in the argument contradict one another. It is this dissonance in his own position which usually provokes his affective responses, rather than his particular commitment to one position or another, or the disagreement between himself and teacher, although the impact of the latter cannot be discounted.

Research in Teaching Style: Experiment One

In the course of our attempts to define operationally RA and SA teaching, we have developed two instruments for the analysis of controversial discussions. Of these two content analysis instruments -- or observational instruments²⁷ -- the first was directed specifically at the problem of differentiating between Recitation and Socratic teaching styles. A second instrument was directed at the problem of evaluating the intellectual process being used by students to deal with a controversial political question after treatment by the different styles. The nature and use of this latter instrument is described in Chapters Ten and Eleven. We are presently developing and will report preliminary results with a third instrument which is directed at a comprehensive description of teacher style, and which has sufficient detail in the cognitive area to be used as an instrument for the evaluation of learning. It includes the major features of both instruments mentioned above, as well as additional dimensions treated earlier in this chapter.

Our first systematic experiment in teaching style was set up to test some initial ideas about distinctions in the teaching styles which we planned on using as a basis for more extensive experimental teaching. In this experiment we asked two major questions:

1. Is it possible for individual teachers to manipulate their behavior and play either a Socratic or recitation role when called upon to do so, or must we identify teachers who have natural Socratic and recitation styles?
2. If it is possible for teachers to manipulate their general style of teaching, what variations occur within a single style, and are there overlapping areas in which one style cannot be differentiated from another?

The instrument developed primarily to differentiate recitation from Socratic teaching was used in an attempt to answer these questions. A brief description²⁸ of it is called for before discussing the experimental results.

Teaching Style Content-Analysis System: Form A. The central problem in developing an observational system for the quantified description of teacher behavior such as that in which we are interested, is to develop a set of categories which will have meaning in terms of the specific teaching styles to which it is to be applied. The instrument, i.e., Form A, which we are

²⁷It should be noted that we will use "observational system," "system of observation," and "observational instrument" interchangeably.

²⁸For a more extensive and detailed treatment of the thought immediately behind this instrument and of the instrument itself, the reader is referred to James P. Shaver, "A Study of Teaching Style."

discussing here actually is composed of two sets of categories.²⁹ Categories 1 through 6 (as shown in Table 12.1) are affective categories; categories 7 through 12 are cognitive categories; and categories 13 and 14 are procedural. In scoring, then, the observer must infer the cognitive significance of verbal behaviors, as well as their affective implications. Each scorable act is categorized in a cognitive category and in an affective category -- including the categories which indicate that either a cognitive or affective message was not discernable. For scoring purposes, the cognitive and procedural categories are grouped together so that no double scoring takes place between them (or it would be a triple scoring scheme).

Our initial attempt to construct an observational system involved a set of affective categories identical with Bales. After several revisions involving both the number and definition of categories, we returned to our starting point. This would seem to confirm, at least in regard to these categories, Bales' conclusion concerning the "essential simplicity and generality of the form as it now stands."³⁰ We did redefine the categories in terms of the classroom situation, drawing examples from that context. We also renamed two of the categories to avoid ambiguities which we found bothersome with a dual classification system. That is, while the labels "solidarity," "tension," "tension release," and "antagonism" all seem clearly affective, Bales' terms "agreement" and "disagreement" have both cognitive and affective connotations. To make these purely affective categories, we substituted the terms "low positive affect" and "low negative affect."

The primary function of the cognitive categories is to answer questions about differences between the cognitive messages expressed in discussions led by teachers using the two teacher styles. Central to the description of the two styles is the extent to which teachers using each deal with descriptions of the controversial case at either the specific or analytic levels as opposed to dealing with the evaluation of events presented in the cases. Categories 8 and 9 are set up specifically to identify differences of this kind. Category 7 (suggests inconsistency) is meant to identify the attempt by the teacher to arouse personal value conflicts on the part of the student by suggesting that he is making contradictory decisions in similar situations (the case and an analogy). Category 12 (analogy) thus has obvious significance. Categories 10 and 11 were employed to make the cognitive subsystem exhaustive.

The procedural categories provide a meaningful exhaustive category system which allows us to gather data about possible procedural problems incidental to the styles. They may also be useful at some later point when and if we compare teaching styles which differ mainly in the procedural area. It should also be pointed out that although we may think the behavior of teachers who are role playing recitation-analytic and Socratic-analytic teaching differ significantly, we will not know this until the assumption has been tested by some such scheme as we are here describing.

³⁰Robert F. Bales, Interaction Process Analysis, p. viii.

TABLE 12.1

BRIEF DEFINITIONS OF THE CATEGORIES IN A PRELIMINARY OBSERVATIONAL
SYSTEM FOR THE DESCRIPTION OF TEACHER STYLE

Affective or Socio-emotional Categories

1. SOLIDARITY -- Status raising language or tone of voice; strong approval or acceptance of another person. Often indicated by enthusiastic acceptance of another's ideas.
2. LOW POSITIVE AFFECT -- Signs of mild approval or acceptance of another person, or of his ideas.
3. TENSION RELEASE -- Action interpreted as tension reducing or attempting tension reduction, e.g., laughing or telling a joke.
4. TENSION -- Behavior indicative of a state of tension, such as stuttering or becoming tongue-tied.
5. LOW NEGATIVE AFFECT -- Statements or acts indicating mild disapproval or rejection of another person, e.g., disbelief of, skepticism about, a statement by the other person.
6. ANTAGONISM -- Deflating, derogatory, or highly negative statements or actions.

NEUTRAL -- Acts or statements with no affective message discriminable by the observer.

Cognitive Categories

7. SUGGESTS INCONSISTENCY -- An attempt to lead another person to see inconsistencies in his values, claims, or definitions.
8. DESCRIPTIVE -- Statements which describe events, i.e., make claims about what reality is like, was like, or will be like.
9. EVALUATIVE -- Statements which evaluate events, i.e., statements of like or dislike, right or wrong, good or bad.
10. REPEATS, SUMMARIZES, FOCUSES -- Statements that restate what has happened during the discussion, or bring attention to what is happening or going to happen.

TABLE 12.1 -- (Continued)

-
11. CLARIFICATION -- Statements that attempt to clarify the content of the discussion, i.e., clear up the meanings of statements or specific words.
 12. ANALOGY -- Statements setting up for consideration a situation similar to the one under discussion. The situation set up may be hypothetical or one which it is claimed existed, exists, or will exist.

NON-COGNITIVE -- Acts with no cognitive -- including procedural -- message discriminable by the observer.

Procedural Categories

13. DIRECTS TASK-ORIENTED BEHAVIOR -- Statements directed at controlling behavior which is in line with the task of the group, or at delineating what that behavior will be.
 14. CONTROLS DEVIANT BEHAVIOR -- Statements directed at controlling behavior which detracts from the accomplishment of the group task.
-

It should also be mentioned that unitization, i.e., the breaking of behavior into bits to be categorized, was based on the "single item of thought." In general, the unit to be categorized was the simple sentence, with complex sentences usually scored as individual units, but with compound sentences broken into their component parts.

Also, symbols were used within each cognitive category to indicate whether the speaker was giving, asking for, or disagreeing with a statement.

Observer reliability. The concept of reliability has already been discussed at some length in Chapter Ten. We might repeat here that there is an ambiguity which should be clarified at the outset. Reliability can refer to the consistency of behavior demonstrated by the person observed, as, for example, referring to a milkman or a paper boy as reliable because he arrives at the same time each day. It can also refer to the consistency between two observations of the same event. It is in this second sense that we are using the term here. Will two observers, both looking at the same specimen of interaction, unitize and categorize the units in a similar fashion?

Two statistical techniques, Chi-Square and Binomial Probability Paper, were used to assess the reliability of observers using the instrument described here. Both methods are suggested and applied by Bales. Probably the statistic most often used to estimate reliability is the Pearson correlation coefficient. As Bales points out, however, "r tends to be relatively insensitive to variations in values with small densities. It is determined mainly by the large values of the distribution."³¹ As with Bales' interaction process analysis, the present scheme produces widely varying category frequencies. For example, with a total of 300 interacts for any discussion, the range of category frequencies may be as great as zero to 200. In comparison to r, Chi-Square "tends to be very much more sensitive to the variations in the pairs of values of smaller magnitude."³²

It should be pointed out, perhaps, that another common method for estimating observer reliability is to compute the percentage of agreement between observers' categorizations. Analysis of variance has also been used for this purpose, and is especially appropriate if the experimenter is interested in analyzing observer agreement along with other components as one of the sources of measurement error, and if more than two sets of scores are available for each observer.

³¹Ibid., p. 102.

³²Ibid., p. 102. Bales also points out that not only is Chi-Square easier than r to compute, but "whereas r is insensitive to the number of acts within categories so long as the proportion of acts within categories to the total acts is constant," Chi-Square provides a "concomitant" estimate of the reliability of unitization and placement in categories, (p. 103). For the reader who is interested, Bales also discusses how the use of Chi-Square he suggests deviates from the usual model, (p. 103).

Reliability attained. Two observers worked together for about four weeks for periods ranging from one to four hours a day while learning to use the instrument. Then, as a basis of agreement was reached, discussions were scored independently for ten-minute periods. The reliability of scoring for these periods was checked using binomial probability paper until suitable reliability was attained.

Bale's adaptation of Chi-Square was used to estimate reliability for the study presented below and the values are presented in Table 12.2. All of the Chi-Square values are well below the maximum allowable to meet the criteria we adopted -- a probability of .50.

Discussions for Experiment One. Using, then, the observational instrument which has been described, an experiment was carried out to determine whether the same teachers could use two styles as different as the Recitation-Analytic and the Socratic-Analytic. This also amounts to an initial attempt to confirm the independent variable of teacher style in the larger experiment described in Chapter Eleven.³³

First the basic concepts to be used in the analysis of societal issues were presented to the seventh grade students. Then a series of 14 cases was used to begin the teaching of the process by which the analytic concepts were to be applied to concrete problems. Each case was to serve as the basis for a one-period discussion and was not embedded within a total unit framework -- except that each case had in common a problem in civil rights as its focus. It was during this period that the Project teachers first attempted a sustained treatment according to the two different teaching styles. (See summary of experimental curriculum in Chapter Nine.)

The ten class periods were spent in the discussion of cases based upon U. S. Supreme Court cases presenting instances in which decisions had to be made in regard to the application of basic principles of American government -- freedom of speech, religious freedom, due process, etc. Each case presented a specific situation involving a conflict between at least two principles. The cases were very brief, two to four double-spaced pages, and were designed as a simple vehicle by which the student could be introduced to the application of the concepts and operations of critical thought to situations of societal conflict.

The teachers were instructed to use at least the first twenty to twenty-five minutes of each class period leading a discussion using the style

³³Actually, another preliminary study, a pilot study in the usual sense, preceded this one. Its purpose was to determine whether observers could learn to use the observational instrument and whether the quantitative data obtained with the instrument would reflect differences in discussions when they were selected on the basis of intuitive conformity to the two models of teaching. Both questions were answered in the affirmative. This study is reported in James P. Shaver, "A Study of Teaching Style."

TABLE 12.2

**SUMMARY OF PROBABILITIES OF CHI-SQUARES RESULTING FROM
ESTIMATION OF OBSERVER RELIABILITY FOR FORTY-ONE
DISCUSSIONS SCORED BY TWO OBSERVERS**

<u>Probability</u>		<u>Socratic</u>			<u>Recitation</u>			<u>Row</u> <u>Total</u>
<u>Less</u> <u>Than</u>	<u>Greater</u> <u>Than</u>	<u>Categories</u>		<u>Total</u>	<u>Categories</u>		<u>Total</u>	
		<u>Aff.</u>	<u>Cog.</u>			<u>Aff.</u>		<u>Cog.</u>
1.00	.90	9	12	21	11	14	25	46
.90	.80	4	6	10	6	5	11	21
.80	.70	2	2	4	2	1	3	7
.70	.50	<u>6</u>	<u>1</u>	7	<u>1</u>	<u>none</u>	1	8
		21	21		20	20		

assigned to that particular group. These first twenty to twenty-five minutes were tape recorded for later scoring with the Form A instrument. In line with the objectives of the Project, this initial discussion was to lead into an analysis of the cases in an attempt to get the students to apply critical thinking skills to the problem under discussion. The teachers were thoroughly familiar with the teaching objectives and the conceptual framework, and all had practiced both teaching roles. In addition, some time had been spent in discussions among the teachers trying to anticipate particular stylistic problems which might arise. In this particular experiment, we were interested only in measuring the initial socratic phase of the Socratic-Analytic style as compared with Recitation-Analytic.

Style confirmation. In scoring the tape recordings, we were first interested in determining whether or not there were over-all differences in behavior -- that is, differences that held up when all categories were compared at once -- when the teachers were using the two styles. For each of the four experimental teachers, we sampled six discussion in which the teacher was trying to use each of the two styles. This was a total of twelve discussions for each teacher. (Because of varying rates of interaction, and some variation in length of scoring time, the frequencies of interests in each category were converted to proportions so as to provide a common base for comparison.)

The basic question at this point, then, is: How are style differences as interpreted and played by the teachers reflected in the over-all quantitative description of the teachers' behavior obtained with the observational instrument? Did the teachers in fact succeed in playing the two different styles (at the abstract level of our categories)? In order to answer this question, we used a statistical technique -- analysis of variance -- which provides an estimate of the extent to which different patterns of categorization are present from one style to the other.³⁴ The results are presented in Tables 12.3 and 12.4. Table 12.3 was computed by pooling data for all teachers, while the computations for Table 12.4 were broken down by teacher. In each case, the ratio testing the interaction between style and category is significant beyond the .001 level.³⁵ We can, therefore, conclude with confidence that whether we look at the teachers as a group or individually, the

³⁴Professor John B. Carroll of the Harvard Graduate School of Education first suggested the possibility of applying a two-way analysis of variance model to our research question about over-all stylistic differences. We classified interacts by style and by category and tested the statistical interaction between style and category.

³⁵For readers who are not familiar with probability and levels of statistical significance, it should be pointed out that this means that the probability of obtaining an F Ratio of this size or larger is less than one in a thousand (indicated on Tables 12.3 and 12.4 by $P=.001$). Another way of saying this is that we would obtain an F Ratio of this size or larger by chance less than one in every thousand times. So we can say that the probability of obtaining an F Ratio of this size by chance is less than one in one thousand or .001 ($P=.001$).

TABLE 12.3

TESTS OF INTERACTION BETWEEN CATEGORY AND STYLE FOR
TRANSFORMED PROPORTIONS**OF TEACHERS' ACTS

Categories	Mean Square		d.f.	Ratio
	R x C	Remainder		
Affective	745.654	11.304	6/322	65.99*
Cognitive	259.894	5.131	13/644	50.66*

*p .001

**In using parametric statistics the researcher always faces the problem of meeting the assumptions which underly the particular model to be applied. Analysis of variance, and the t test which will be reported in following tables, assume that the data come from normally distributed populations and that the variances are homogeneous. These assumptions become untenable when results are expressed in proportions. One way of meeting this problem is to transform the proportions to angles using an inverse sine transformation, and then use these figures as a basis for computation. This is the procedure we followed. For those who would like to look at the following tables in terms of proportions, Table 12.11 provides the information for quick transformation. [For discussions of the assumptions mentioned above and of the appropriateness of using the inverse sine transformation, the interested reader is referred to the following sources: Virginia L. Senders, Measurement and Statistics (New York: Oxford University Press, 1958), p. 498; Helen M. Walker and Joseph Lev, Statistical Inference (New York: Henry Holt & Co., 1953), p. 424; Quinn McNemar, Psychological Statistics (New York: John Wiley and Sons, 1955), p. 255; and Allen Lewis Edwards, Experimental Design in Psychological Research (New York: Rinehart and Winston, 1950), p. 203.]

TABLE 12.4

TESTS OF INTERACTION BETWEEN CATEGORY AND STYLE FOR TRANSFORMED
PROPORTIONS OF ACTS OF INDIVIDUAL TEACHERS

Teacher	Categories	Mean Square		d.f.	Ratio
		R x C	Remainder		
B	Affective	402.841	16.950	5/60	23.77*
	Cognitive	154.975	5.080	9/100	30.51*
G	Affective	185.825	13.132	5/60	14.15*
	Cognitive	49.200	3.300	9/100	14.91*
O	Affective	275.538	10.699	5/60	25.75*
	Cognitive	228.292	6.082	9/100	37.54*
S	Affective	113.829	7.434	5/60	15.31*
	Cognitive	24.546	4.744	9/100	5.17*

* $p < .001$

categories tend to contain different mean values of the transformed proportions depending upon the style of teaching which is to be in use. It should be noted that the F Ratios are very sizable, indicating great differences in teaching style as observed.

With the conclusion that the teachers were able to effect great over-all differences in style, a second question becomes pertinent: Which of the individual categories contribute to these global differences? The following generally stated hypotheses, in conjunction with Table. 12.5, suggest the categories in which we would expect to find stylistic differences particularly reflected.

1. In the area of affect we would anticipate that while recitation teaching is going to require more low positive affect in carrying out its general program of reinforcement of correct responses, there is no reason to expect more highly loaded positive affect, or solidarity. Moreover, the highly argumentative socratic situation should reflect a greater amount of both low negative affect and antagonism than the recitation. And we expected the teachers to use more tension release -- telling of jokes and laughing -- with the socratic style in order to keep tension under control. The affect-laden socratic situation also would lead to a prediction of fewer affectively neutral acts.
2. Because of the differing focuses of discussion with the two styles, the teachers should make more descriptive statements with the recitation style, and ask for more evaluations with the socratic style. It is presumed that socratic teaching will emphasize consideration of inconsistencies in the student's position, and therefore the teachers will suggest more evaluative inconsistencies with this style while avoiding a statement of their own positions. While the argumentative situation might have indicated such a result, we did not anticipate that socratic teaching would necessitate any greater amount of disagreement with descriptive statements made by students.
3. Two categories seem rather directly related to the different nature of lesson organization for the two teaching styles. While socratic teaching involves an intensive and lengthy intercourse with individual students centered around the values expressed by the student, recitation teaching requires planning organized around the teacher's conception of the adequate treatment of the problem. It is necessary, as recitation moves from one student to another and as each responds to the teacher's queries, that this organizational structure be maintained. We therefore would expect that the recitation teaching would result in a greater proportion of acts in the repeats, summarizes, and focuses category. This same factor might well have a similar effect in the clarification category.
4. It is obvious from our stylistic models that socratic teaching should result in many more analogies than would be the case with recitation teaching.

TABLE 12.5

PREDICTED DIFFERENCES IN CATEGORIZATIONS OF BEHAVIOR
FOR TEACHERS ATTEMPTING TO USE THE TWO STYLES

CATEGORY*	STYLE FOR WHICH GREATER PROPORTION PREDICTED	FATE OF HYPOTHESIS
Solidarity	No Difference	Rejected
Low Positive	Recitation	Accepted
Tension Release	Socratic	Rejected
Low Negative	Socratic	Accepted
Antagonism	Socratic	Accepted
Neutral	Recitation	Accepted
Evaluative, Gives	No Difference	Rejected
Evaluative, Asks For	Socratic	Accepted
Suggests Inconsistency, Evaluative	Socratic	Accepted
Descriptive	Recitation	Accepted
Disagreement with Description	No Difference	Rejected
Repeat, Summarize, Focus	Recitation	Accepted
Clarification	Recitation	Rejected
Analogy	Socratic	Accepted
Directs Task-Oriented Behavior	Recitation	Accepted

*Categories are not included in which there were insufficient frequencies for analysis or which were not considered central to the styles.

5. A high frequency behavior categorized in the category of direct task-oriented behavior is calling on students. Because of the difference in duration of periods of student-teacher interaction, we can expect a greater proportion of categorizations in this category when the recitation style is to be in use.

Table 12.5 indicates the fate of hypotheses for specific categories, and Table 12.6 presents the actual results of computing t values to test the differences in transformed proportions for the categories. Ten of the sixteen predictions were confirmed. A brief comment is in order on those which were rejected. While the teachers did use more behavior categorized as solidarity when using the recitation style, this seems to be a reflection of the reinforcing function. But this result raises the question of the appropriateness of timing of that solidarity and tension release evidenced during Socratic teaching. The greater proportion of evaluative statements during Socratic teaching probably reflects two factors: the difficulty of talking about value issues without expressing one's own commitments; and, the expression of one's own values in order to draw out value statements on the part of students. It should be noted that the difference is between about one percent on the one hand and two and one-half on the other. That the teachers disagreed with more descriptive student statements while using the Socratic style is probably a function of the argumentative nature of Socratic discussions, as well as a reflection of the success of recitational programming -- i.e., setting up a series of questions based on assigned material and sequenced so that student errors in response are avoided. The lack of difference in clarification simply indicates that the discussion needs to be clarified at certain points with either style.

Variation of behavior within teaching styles. Despite some discrepancies in individual categories, it seems safe to conclude that our teachers were able to manipulate their general style of teaching. This, then, raises a second major question: Did significant variations in behavior occur within either style, and if so, are these variations so great as to result in overlapping areas in which one style cannot be discriminated from another on the basis of category proportions even though there are significant differences in proportions when styles are compared with teachers pooled?

Table 12.7 indicates the results of classifying interacts on the basis of category and teacher for each style, and then testing the significance of interaction using once again a two-way analysis of variance model. This will tell us whether there are different patterns of behavior -- i.e., categorization of behavior -- from teacher to teacher when they are all presumably using the same style. The results would indicate that this is so, at levels of significance ranging from .05 to .001.

Tables 12.8 and 12.9 present the results of an effort to explore which categories account for these within-style differences. Here a simple, or one-way, analysis of variance was used to test the significance of differences among multiple means. Considering educational claims about the impact of teacher personality and immediate situational factors upon teacher behavior,

TABLE 12.6

COMPARISON OF MEANS OF TRANSFORMED PROPORTIONS OF
INTERACTS FOR TEACHERS ATTEMPTING TO USE THE TWO STYLES

Category	Style		Difference	t*
	Socratic	Imitation		
Solidarity	.641	1.405	.765	4.35****
Low Positive	4.576	9.539	4.963	10.28**
Tension Release	.495	.465	.030	.25
Low Negative	10.517	2.326	8.192	34.42**
Antagonism	8.879	.151	8.729	6.32**
Neutral	34.624	48.970	14.346	7.82**
Suggests Evaluative				
Inconsistency	5.492	.161	5.331	7.50**
Descriptive	21.231	34.930	13.699	7.66**
Specific Disagreement	.904	.489	.415	2.40***
Evaluative	1.484	.500	.983	5.22
Evaluative, Asks For	8.239	3.005	5.154	7.15**
Repeats, Summarizes,				
Focuses	5.512	7.770	2.258	4.75**
Clarification	2.064	2.336	.272	1.01
Analogy	6.871	1.362	5.508	5.84**
Directs	4.374	5.879	2.505	6.77**
Non-Cognitive	1.456	1.672	.216	1.15

* d.f. = 23 for all computations

** p (.001 one-tailed test

*** p (.05 two-tailed test

**** p (.001 two-tailed test

TABLE 12.7

INTERACTION BETWEEN CATEGORY AND TEACHER FOR TRANSFORMED
PROPORTIONS OF ACTS OF TEACHERS ATTEMPTING TO USE THE TWO STYLES

Style	Categories	Mean Square		d.f.	Ratio
		R x C	Remainder		
Socratic	Affective	42.218	20.172	12/120	2.09*
	Cognitive	36.646	6.298	27/200	5.82***
Recitation	Affective	7.350	3.101	15/120	2.37**
	Cognitive	8.578	3.304	27/200	2.50***

* $p < .05$ ** $p < .01$ *** $p < .001$

TABLE 12.C

COMPARISON OF MEANS OF TRANSFORMED PROPORTIONS OF INTERACTS
FOR TEACHERS ATTEMPTING TO USE THE SOCRATIC STYLE

CATEGORY	TEACHER				F RATIO*
	B	G	O	S	
Solidarity	.44	.69	.24	1.19	5.09***
Low Positive	3.59	5.03	4.09	5.62	.194
Tension Release	.31	.27	1.03	.36	4.17**
Low Negative	10.18	8.10	14.09	9.69	4.25**
Antagonism	14.40	9.02	5.75	5.34	2.50
Neutral	30.52	36.86	33.14	37.97	-----****
Evaluative	1.68	1.83	1.03	1.39	1.20
Evaluative, Asks for Suggests Evaluative	10.31	5.75	8.97	6.93	2.44
Inconsistency	7.59	3.94	8.21	2.53	7.03***
Descriptive	16.57	25.81	15.03	26.51	7.73***
Descriptive, Disagreement	.52	.83	1.20	1.08	1.32
Repeats, Summarizes, Focuses	4.95	7.15	3.76	6.21	5.34***
Clarification	1.31	2.57	2.30	2.00	4.70**
Analogy	9.88	4.56	8.56	4.48	2.61
Directs Task- Oriented Behavior	3.59	3.50	5.50	4.80	3.31**

* d.f. = 3/20 for all categories

** p. < .05

*** p. < .01

**** Ratio not computed because within rows variance larger than
between rows variance.

TABLE 12.9

COMPARISON OF MEANS OF TRANSFORMED PROPORTIONS OF INTERACTS
FOR TEACHERS ATTEMPTING TO USE THE RECITATION STYLE

CATEGORY	B	TEACHER			F RATIO*
		G	O	S	
Solidarity	.61	1.42	1.02	2.56	6.02***
Low Positive	8.51	8.39	10.39	10.87	2.55
Tension Release	.81	.28	.70	.08	3.50**
Low Negative	2.76	2.57	2.11	1.87	-----****
Antagonism	.18	.07	.09	.26	1.14
Neutral	50.32	50.61	46.39	46.56	1.84
Evaluative	.43	.72	.51	.34	3.77**
Evaluative, Asks for	3.31	3.26	1.81	3.96	1.39
Suggests Evaluative					
Inconsistency	.06	.15	.34	.11	-----****
Descriptive	32.67	36.12	38.67	32.25	2.63
Descriptive,					
Disagreement	.55	.40	.56	.44	-----****
Repeats, Summarizes,					
Focuses	8.57	8.28	5.43	7.80	2.81
Clarification	2.15	2.36	2.76	2.08	-----****
Analogy	1.56	1.52	1.36	1.01	-----****
Directs Task-					
Oriented Behavior	7.61	5.76	6.28	7.88	4.01**

* d.f. = 3/20 for all categories

** p. .05

*** p. .01

****Ratio not computed because within rows variance larger than between rows variance.

it is not surprising to find that with both styles there is considerable variation among the teachers on several categories. To be specific, differences significant beyond at least the .05 level are indicated for nine categories when the role expected was Socratic teaching, and for five categories when the teachers were playing the recitation role.

We must then ask to what extent these significant differences within style result in overlap between styles. Table 12.10 was set up with this question in mind. In two instances in categories for which differences were predicted between styles and which reflected differences within style, the differences between means most extreme from the predicted directions were not significant at the .05 level. In other words, differences of that magnitude would occur by chance more than five times out of a hundred, so we do not consider the means as different from one another.

Little needs to be added to the discussion of results as they have been presented. One point might be reiterated and one additional point made, however. First, the size of the F Ratios which resulted when differences in behavior between styles were tested should be re-emphasized. They undoubtedly reflect the differences in behavior which at least some teachers can produce when it is their intention to do so, and they do indicate that in some major respects groups of students did receive different systematic treatment. The potential utility of the analysis of within-style differences which can be pursued when a system of observation is used must also be stressed. Considerable increment in information beyond intuitive judgments of style is gained through systematic observation and can be used in explaining learning outcomes. This is clearly evident when one looks at those tables which describe the differences and similarities. It is the relationship of specific dimensions of behavior, as measured through such systematic observation, to learning outcomes that we think holds particular promise for teacher effectiveness research in the future.

Research in Teaching Style: Experiment Two

As a result of Experiment One we gained some confidence in the fact that the teachers involved in the Project were able to play the Socratic and Recitation styles effectively, and they appeared to be able to manipulate the styles with some degree of consistency. We were next interested in the following questions:

- a) Is the Socratic-Analytic style more effective than the Recitation-Analytic style in teaching the student to understand and handle problems involving political controversy?
- b) If the two styles did have differing effects on learning outcomes, would the RA style more efficiently teach the analysis of abstract controversial dialogues, and the SA style teach more effectively the ability to handle a live dialogue in which the student himself is actually required to defend a position?

TABLE 12.10

DISTRIBUTIONS OF MEANS OF TRANSFORMED PROPORTIONS FOR CATEGORIES FOR WHICH DIFFERENCES WERE PREDICTED BETWEEN STYLES AND IN WHICH THERE WERE SIGNIFICANT DIFFERENCES AMONG TEACHERS WITHIN EITHER THE SOCRATIC OR RECITATION STYLE

LOW NEGATIVE				SUGGESTS EVALUATIVE INCONSISTENCY			
Teacher Socratic		Teacher Recitation		Teacher Socratic		Teacher Recitation	
O	14.09	B	2.76	O	8.21	O	.34
B	10.18	G	2.57	B	7.59	G	.15
S	9.69	O	2.11	G	3.64	S	.11
G	8.10	S	1.87	S	3.53	B	.06

DESCRIPTIVE*				REPEAT, SUMMARIZE, FOCUS			
Teacher Socratic		Teacher Recitation		Teacher Socratic		Teacher Recitation	
S	26.51	O	38.67	G	7.16	S	8.57
G	25.81	G	36.12	S	6.21	G	8.28
B	16.57	B	32.67	B	4.95	S	7.80
O	16.03	S	32.25	O	3.78	O	6.43

DIRECTS TASK-ORIENTED BEHAVIOR*			
Teacher Socratic		Teacher Recitation	
O	5.50	S	7.88
S	4.80	B	7.61
B	3.69	O	6.28
G	3.50	G	5.76

*Difference between the two means each of which is most extreme from the predicted direction for each style is not significant at the .05 level, t test.

TABLE 12.11

VALUES RESULTING FROM THE TRANSFORMATION OF SOME PROPORTIONS
WITH THE INVERSE SINE TRANSFORMATION*

Proportion	Degrees	Proportion	Degrees
.01	.567	.25	14.483
.02	1.150	.30	17.450
.03	1.717	.35	20.483
.04	2.300	.40	23.583
.05	2.867	.45	26.750
.06	3.433	.50	30.000
.07	4.017	.60	36.867
.08	4.583	.70	44.433
.09	5.167	.80	53.133
.10	5.733	.90	64.150
.15	8.633	1.00	90.000
.20	11.533		

*This table is included to assist readers who wish to estimate the proportions represented by the transformed figures presented in the preceding tables.

A basic design was set up within the experimental school to investigate these questions. Each regular section composed of 23 to 27 students was divided into two groups. Students were taught approximately half the time in these subgroups, and were always split when they were to discuss a controversial case. There were some 50 controversial cases used over the two year period. (See Chapter Nine.) The split groups were matched as closely as possible on the following variables: sex, intelligence test scores in school records, scores on the Wagnis Test, scores on the ITED No. 5, and oral participation rate of individual students as determined by the ratings of the experimental teachers. The division of each section was not made by blind pairing of scores alone. Some effort was made to have the four experimental teachers feel intuitively that each experimental group or subsection had a reasonably equal counterpart in the other half of the section.

By dividing the five sections in half, ten groups were taught over a two year period on some 50 cases according to systematic style differences. With four teachers available, two teachers taught two groups and two teachers taught three groups. Since the five regular sections are grouped by the school administration on the basis of general scholastic ability, there was one very able section (two subsections), two sections of above average ability (four subsections), and two sections of students with average or below average ability (four subsections).

Ideally, our research design should have been set up to investigate not only the effects of teacher style, but also interaction effects between style and teacher, style and student ability, and teacher and student ability. Unfortunately we had neither the number of groups nor the number of teachers to set up such an elegant design. Moreover, one of the teachers who was a regular member of the school faculty had duties outside the Project which further limited the flexibility of our research design.

TABLE 12.12

TYPES OF GROUPS TAUGHT AND STYLES USED BY
FOUR EXPERIMENTAL TEACHERS

ABILITY OF GROUP	TEACHER			
	B	G	O	S
High	-----	Soc (6)*	Rec (6)	-----
High Average	Rec (7) Soc (4)	Rec (4)	-----	Soc (7)
Average	-----	-----	-----	-----
Low Average	Soc (5)	-----	Soc (3)	Rec (3) Rec (5)
Low	-----	-----	-----	-----

*Numbers are designations of specific regular class groups. Thus, group 6 is a high ability group split into a recitation and Socratic subsection, etc.

Rec = Recitation Style

Soc = Socratic Style

TABLE 12.13

SELECTED VARIABLES DESCRIBING TEN MATCHED GROUPS

Groups	N	No. Boys	Ited No. 5		Wagm1s		I.Q.	
			Mean	S.D.	Mean	S.D.	Mean	S.D.
R-6*	13	5	28.08	7.40	44.50	7.25	124.67	6.90
S-6	12	6	27.17	6.09	42.00	7.98	126.91	9.64
R-4	13	7	23.83	7.54	37.83	6.11	116.83	8.32
S-4	13	5	23.46	4.50	39.53	4.52	116.00	8.71
R-7	11	4	25.00	3.81	37.64	5.20	112.80	7.25
S-7	12	8	25.33	7.65	35.69	7.20	114.15	8.05
R-3	13	8	16.11	3.17	32.44	4.65	105.11	7.42
S-3	11	7	16.50	4.49	30.92	4.72	105.00	8.79
R-5	13	8	16.25	5.13	28.58	6.40	101.81	12.76
S-5	13	6	18.30	4.22	33.20	5.31	99.50	12.21

*Numbers correspond to types of groups taught as in Table 12.12.

R=Recitation

S=Socratic

The characteristics of the groups actually available, the teaching assignments, and the relationships between groups, styles and teachers are summarized in Tables 12.12 and 12.13.

Results of the comparison between recitation and socratic groups. The results of the comparison of the two treatments called Recitation-Analytic and Socratic-Analytic (or simply recitation and socratic) are summarized in Table 12.14. The tests presented in Table 12.14 have all been described, except the "Unit Tests." This is a total score obtained from four teacher-made tests given at the end of the four major experimental substantive units. (The four units were on school desegregation, fair business competition, labor, and the New Deal.) Each test contained a section demanding simple recall of factual material, a section requiring the student to apply the analytic concepts to a particular aspect of the topic of the unit, and one or two open-ended essay questions. While the tests were constructed and scored under the pressure of regular school business, they contained so many items that the total score is probably an extremely reliable index of the student's immediate receptivity to the treatment provided in the experimental program.

The most obvious fact about the results of the differential treatment experiment is that groups taught by the two treatments turn out to behave in an astonishingly similar way on all measures of learning outcomes administered. This result is consistent with general findings in "methods" research. Few experiments, however, have operated under such carefully controlled curricula, or have actually demonstrated marked differences in teaching style quantitatively, as we have. Regarding the general finding we might make the following comments.

1. The instruction was probably generally very effective across styles. All teachers were selected not only because of their ability to play to style roles well, but also because of their general excellence as teachers. The materials were especially prepared to teach a particular analytic framework. The teachers were involved in the development of the materials, the preparation of lessons based on the materials, as well as the actual teaching. With this combination of resources, it is reasonable to assume that a powerful instructional situation resulted, regardless of the particular style employed by the teacher.
2. All measures except SIAT No. 4 (the Socratic Discussion Test) assessed the student's use of analytic concepts either implicitly in the substance of an argument or explicitly in a non-stress situation. It is therefore surprising, actually, that the recitation groups did not perform consistently better on all measures other than SIAT No. 4. Since the analytic framework was learned as well by the socratic groups, one might assume a tremendous redundancy of treatment, so that despite much less explicit discussion of the framework in the socratic groups, they grasped certain basic elements of the system as well as did the recitation groups. It is very difficult to

TABLE 12.14

COMPARISON OF EXPERIMENTAL GROUPS TAUGHT BY
SOCRATIC AND RECITATION STYLES DURING DISCUSSIONS OF CONTROVERSIAL CASES

Test	Recitation			Socratic			t
	N	Mean	S.D.	N	Mean	S.D.	
<u>Content Tests:</u>							
Ited No. 1	53	35.38	10.49	55	36.04	9.66	$\sqrt{F} = 0.03^*$ 0.35
U.S. History	52	110.88	32.62	55	109.46	28.43	
Principles of American Citizenship	52	33.40	8.15	56	35.00	7.67	1.04
<u>General Reasoning Tests:</u>							
Ited No. 5	55	32.73	10.10	52	32.93	10.36	0.10
Wagnis	53	44.25	9.82	56	45.61	9.17	0.74
<u>Interest Tests:</u>							
Headlines							
Project Topics	55	24.00	3.79	47	23.87	4.91	0.14
General Topics	55	26.85	4.05	47	26.49	4.05	0.44
Unit Test	56	161.76	46.42	53	160.39	44.22	0.16
<u>Project Political Analysis Tests:</u>							
SIAT No. 1	53	12.02	4.84	55	12.71	3.87	0.82
SIAT No. 2	56	39.39	8.88	52	39.23	8.95	0.09
SIAT No. 3	53	29.34	11.03	53	32.19	10.50	$\sqrt{F} = 0.52^*$ 0.30
SIAT No. 4	51	19.24	10.40	52	19.83	9.72	

*Covariance used because of significant difference on pretest.

explain, however, why the socratic groups did not do better on SIAT No. 4. Two of the most plausible explanations might be explored. First, there may be something deficient in the test. The most obvious deficiency, we think, is in the amount of behavior sampled (actually only 15 minutes). While this gives us a great deal of behavior when compared with pencil-and-paper tests (some 225 acts), it constitutes a very brief discussion. It also makes the assumption that there is homogeneity across topics which may be used as the basis of such a discussion. This assumption requires further exploration. Second, there may be something deficient in the instructional situation: the students may have been too young to gain substantial benefit from use of the analytic concepts in a persuasive situation; or the treatment may have been too brief to have made a significant impact. It is our impression that both of these situations may, to some extent, be true. We would say that most of the students had difficulty translating the analytic framework into a usable tool when engaged in a dialectical discussion. Whether this difficulty is related to intellectual maturity, or whether it is a general phenomenon is difficult to say.

It is very possible that there is an interaction between individual personality characteristics and the student's ability to benefit from one style or the other. In other words, some students may learn from one style and be inhibited by the other style, and vice versa. This hypothesis has been very much in our thinking since we discovered in an earlier study that while, on the average, groups involved in high threat discussions do not learn any more or less than groups involved in low threat discussions, groups of students who participate in high threat discussions do have more heterogeneous scores after treatment, i.e., there is a significant increase in variance. Resnick also uncovered a similar finding.³⁶ This hypothesis will be explored more fully in Chapter Thirteen when we discuss the general effect of student personality on the learning of controversial issues, and the possible relationship between personality and socratic or recitation teaching.

Comparison of the ten experimental subgroups. While our design does not permit us to determine by statistical tests the complex relationships between the effects of individual teachers, the intellectual quality of the group, and teaching style, we might gain some insight into such relationships by testing whether or not there are significant differences across groups on each of the measures. And if there are, we might look at individual mean

³⁶Lauren B. Resnick, "The Handling of Contradictory Information: A Study in the Effects of Cognitive Need, Cognitive Style, and Motivational Arousal," 1962 (Mimeographed).

scores to see which particular groups account for the differences. These data are presented below in Table 12.15.

The first and most obvious finding apparent from the data in Table 12.15 is that adjusting individual scores for pretest score and I.Q. estimate does not cancel out the effects of general intellectual competence. It is clear that the more competent groups still do better after the covariance adjustment. A very simple way to demonstrate this fact is to rank the ten groups on project developed measures and summarize the relationship between the frequency of ranks above and below the median (the top and bottom five ranks) and the ability of the subgroups in four-fold tables. These results are summarized in Table 12.16.

These tables present rather striking evidence that the probability of doing better as a result of treatment is much better if one is in a high-average or high ability group rather than in a low-average ability group. It should be emphasized that this is not an obvious finding, since, as we pointed out above, we adjusted statistically for the effects of academic aptitude and pretest score before making these comparisons. Two explanations might be offered to interpret this result. First, one might hypothesize that the instructional situation (or the treatment) for high ability groups is necessarily different from that of low ability groups. High ability groups can progress faster; they can deal with material in greater detail and with greater sophistication. Thus the proper correction cannot be made statistically, because the quality of the instructional situation for high and low ability groups is fundamentally different. A second interpretation relates to the confidence one has in measures of general intellectual ability or scholastic aptitude. There may be so much slippage in these measures that they do not take into account certain intellectual qualities which help one benefit from the experimental curriculum. Evidence for this conclusion is found in the relatively high correlation between learning outcome measures and WAGMIS which still remains after partialing out an estimate of I.Q. Regardless of which interpretation one chooses -- and both may be involved in a complex explanation -- this finding has very important implications for future research.

We might note one final point. In exploring these data further we discovered that while there is a clear relationship between the general intellectual ability of the group and growth on the learning outcome measures, there is some consistency regarding which "above average" groups fall below the median. (Of course, in each four-fold table at least one "above average" group must fall below the median.) Of the six instances in which this occurs, four involve Recitation Group 4 and two involve Socratic Group 6. It happens that these two particular groups were taught by the same teacher, G. We scored³⁷ some 20 teaching performances (approximately five for each teacher) in an attempt to assess how G's teaching might differ from the other three

³⁷ We used the interaction system described earlier in this chapter, a scoring manual of which is included in Appendix C.

teachers'. There was, however, surprising similarity of profiles across categories among the four teachers. The only obvious difference in G's teaching was a tendency to deal more in factual generalizations and definitional questions (especially those related to factual questions) and less in value questions.

TABLE 12.15

COMPARISON OF MEANS OF TEN SUBGROUPS TAUGHT BY FOUR TEACHERS
WITH RECITATION AND SOCRATIC STYLES ADJUSTED FOR PRETEST SCORE AND I.Q.

Test	Style	Subgroups					DF	F-ratio
		3 ^a (Average and low ability)	5 ^a	4 ^b (Above average ability)	7 ^b	6 ^a (High ability)		
<u>Content Tests:</u>								
Itd No. 1 ^b	Rec	[Sc N=8 S N=11 G N=12 B N=10 O N=12]						
	Soc	[O N=11 B N=9 B N=12 S N=12 G N=12]					9/103	2.67
U. S. History	Rec	91.15	77.75	116.77	114.72	140.42		
	Soc	92.42	87.60	109.12	119.57	133.19	9/100	6.82
Principles of American Citizen-ship	Rec	35.68	29.61	34.47	32.86	35.62		
	Soc	32.43	30.96	36.25	39.12	34.34	9/103	2.67

^aCode numbers indicate regular class sections, and each includes one recitation and one socratic subgroup.
^bBecause of negligible differences among adjusted means for Ited No. 1 and Headlines, General, the means for these tests are not presented.

C, G, B, and O are code names for project teachers; N=model group for all comparisons in each subgroup.

TABLE 12.15 -- continued

Test	Style	Subgroups				DF	F-ratio
		3 (Average and low ability)	5 (Above average ability)	7 (High ability)	5		
<u>General Reasoning Tests:</u>							
Itd No. 5	Rec	26.91	26.06	34.74	34.34	38.56	
	Soc	26.64	27.63	38.84	32.81	40.13	9/102 4.58
	Rec	40.53	36.73	39.12	46.79	51.17	
Magm's	Rec	40.53	36.73	39.12	46.79	51.17	
	Soc	39.75	38.34	47.51	49.04	50.37	9/104 6.03
<u>Interest Tests:</u>							
Headlines Project Topics	Rec	25.05	22.78	22.59	23.39	26.01	
	Soc	23.38	24.19	22.77	21.88	27.51	9/97 2.12
	Rec						
General Topics	Rec						
	Soc						9/97 1.40
Unit Test	Rec	153.29	129.35	169.01	168.39	187.91	
	Soc	127.36	152.64	176.79	172.77	164.12	9/105 3.87

TABLE 12.15 -- continued

Test	Style	Subgroups				DF	F-ratio
		3 (Average and low ability)	5 (Above average ability)	7 (High ability)	6 (High ability)		
<u>Project Political Analysis Tests:</u>							
SIAT No. 1	Rec	10.27	8.07	11.34	13.77	16.23	
	Soc	9.70	10.34	12.53	14.85	15.11	9/105 6.82
SIAT No. 2	Rec	44.98	43.93	47.23	49.18	50.59	
	Soc	38.78	42.84	50.12	50.09	48.56	9/103 3.91
SIAT No. 3	Rec	27.64	25.08	26.44	33.59	36.65	
	Soc	25.98	25.15	33.75	40.95	29.84	9/101 4.02
SIAT No. 4 ^d	Rec	17.38	20.20	16.64	22.30	19.50	
	Soc	16.60	18.13	19.50	25.45	18.73	9/98 0.84

^dSIAT No. 4 is included despite insignificant differences. The findings are of special interest because of the overlap of scores among groups of differing ability. Unadjusted means are presented because of the low correlation between posttest score and either I.Q. or pretest score.

TABLE 12.16

FOUR-FOLD TABLES DESCRIBING RELATIONSHIP BETWEEN
ABILITY OF EXPERIMENTAL SUBGROUPS AND LEARNING OUTCOMES

a. SIAT No. 1

	Top Five Ranks	Bottom Five Ranks	
High Ability Groups	5	1*	6
Low Ability Groups	0	4	4
	5	5	10
*Recitation Group	4		

b. SIAT No. 2

	Top Five Ranks	Bottom Five Ranks	
High Ability Groups	5	1*	6
Low Ability Groups	0	4	4
	5	5	10
*Recitation Group	4		

c. SIAT No. 3

	Top Five Ranks	Bottom Five Ranks	
High Ability Groups	5	1*	6
Low Ability Groups	0	4	4
	5	5	10
*Recitation Group	4		

TABLE 12.16 -- continued

d. SIAT No. 4

	Top Five Ranks	Bottom Five Ranks	
High Ability Groups	4	2*	6
Low Ability Groups	1	3	4
	5	5	10

*Recitation Group 4 and Socratic Group 6

e. UNIT TESTS

	Top Five Ranks	Bottom Five Ranks	
High Ability Groups	5	1*	6
Low Ability Groups	0	4	4
	5	5	10
*Socratic Group	6		

APPENDIX C

SCORING PROCEDURES FOR AN OBSERVATION SYSTEM
TO DESCRIBE TEACHING STYLE AND LEARNING OUTCOMES

SAMPLE SCORING SHEET

[illegible]

APPENDIX C

SCORING PROCEDURES FOR AN OBSERVATION SYSTEM
TO DESCRIBE TEACHING STYLE AND LEARNING OUTCOMES

A. General Instructions

1. Scoring Statement Posture:
 - a. declarative statements, score 1
 - b. interrogative statements, score 2
 - c. questioning a previous statement, doubting, or self-doubting, score 3
2. Scoring Teacher and Student or Interviewer and Subject:
 - a. circle number if the speaker is teacher or interviewer
 - b. if speaker is student or subject, simply put posture number without circle

a) -	b)	c)	d)																								
<table><tr><td>g</td><td>c</td><td>q</td></tr><tr><td>d</td><td>1</td><td>a</td></tr></table>	g	c	q	d	1	a	<table><tr><td>g</td><td>c</td><td>q</td></tr><tr><td>d</td><td>(1)</td><td>a</td></tr></table>	g	c	q	d	(1)	a	<table><tr><td>g</td><td>c</td><td>q</td></tr><tr><td>d</td><td>2</td><td>a</td></tr></table>	g	c	q	d	2	a	<table><tr><td>g</td><td>c</td><td>q</td></tr><tr><td>d</td><td>(2)</td><td>a</td></tr></table>	g	c	q	d	(2)	a
g	c	q																									
d	1	a																									
g	c	q																									
d	(1)	a																									
g	c	q																									
d	2	a																									
g	c	q																									
d	(2)	a																									

- Thus, a) if student makes a statement, score 1
- b) if teacher makes a statement, score (1)
- c) if student asks a question, score 2
- d) if teacher asks a question, score (2)

3. Scoring Static Operations: To score one of the major categories or static operations place the symbol indicating posture and speaker (teacher or student) beside the operation performed, within the cluster of symbols, in each square.
- *SAMPLE:
- | | | | | | | | |
|--------------|--|---|---|---|---|---|---|
| REPEAT, ETC. | <table border="1"><tr><td>g</td><td>c</td><td>q</td></tr><tr><td>d</td><td>2</td><td>a</td></tr></table> | g | c | q | d | 2 | a |
| g | c | q | | | | | |
| d | 2 | a | | | | | |

4. Scoring Dynamic Operations: To score one of the dynamic operations, place a line through one of the symbols in the scoring square identifying the first letter of the operation.
- SAMPLE:
- | | | | | | | | |
|-----|---|--------------|---|--------------|---|---|---|
| GLC | <table border="1"><tr><td>g</td><td>c</td><td>~~q~~</td></tr><tr><td>d</td><td>1</td><td>a</td></tr></table> | g | c | ~~q~~ | d | 1 | a |
| g | c | ~~q~~ | | | | | |
| d | 1 | a | | | | | |

5. Scoring the framework of persuasion:

Most acts in any discussion will be oriented to the actual solution of a problem within the (usually implicit) framework of commonly accepted categories. That is, statements will either be judgments about the merits of a situation, or evidence, definitions, analogies, etc.,

*Samples for scoring operations are taken from a whole range of categories and are not meant to indicate that the operation occurs only in a specific category

5. (continued)

SAMPLE:

GVJ	g c q
-	d 1 /

used to clarify or extend a judgment. In all these cases the discussion is to be considered substantive and no score recorded.

When a statement is made which tries to define a more abstract structure of the problem so that its substance can be more adequately handled, the ANALYSIS score should be made by marking the a in the score box. The most clearly recognizable aspect of ANALYTIC statements is that they make no direct or indirect recommendations about the solution of the problem. For an example, see discussion in Chapter Twelve.

6. Scoring Discussion Posture --
DIALECTICAL-DESCRIPTIVE:

SAMPLE:

SVJ	g c q
	d 2 a

If the teacher regards the content as problematical, so that answers are not treated as definitely "right" or "wrong," the discussion is to be considered DIALECTICAL, and the d in the scoring box should be marked. Otherwise, the discussion is descriptive, and no mark should be made. While both descriptive and dialectical teaching may assume a problem or question before the class, the latter posture necessarily does. Furthermore, the dialectical posture assumes that the student is permitted a degree of autonomy in the way he deals with the problem. It is assumed that the outcome of discussion will not be the sum of the teacher's and students' contributions, but a result of teacher-student interaction.

In practice, many dialectical statements may appear to be coercive: the teacher may seem to be trying to force the student to agree with him. If this is done to elicit an argument or defense of position by the student, the statement is clearly to be scored DIALECTIC.

Example (Teacher statements):

--"Don't you agree that unions ought to be more strictly regulated?"

--"How can you possibly believe that medicare proposals are anything but socialized medicine?"

B. Unitization Conventions

Basic Rule: Simple and complex sentences are the basic unit.
Compound sentences are scored as two units.

Complex sentences are scored as two units when the dependent clause(s) would be scored in a different category from the main clause. [For example, "Negroes should be given preference in housing because the whites are responsible for bringing them over here in the first place." VALUE JUDGMENT + GENERAL CLAIM.]

A SPECIFIC CLAIM that refers to a single event or instance is scored as one unit, regardless of how many sentences it takes to describe it.

SOURCE is scored as a separate unit, even though it is part of a simple sentence which is scored in another category.

DEFINITION: each criterion of a definition is scored as a separate unit, regardless of whether it takes more or less than a simple sentence. If the same criterion is repeated, it is scored as REPEAT. If several different criteria are given in a row, they are all scored as separate units under DEFINITION. Elaboration of a definition in which no new criteria are added receives one unit.

CLARIFICATION: A "single thought unit" of clarification supercedes the single sentence. Several sentences given as a single point of clarification, then, are scored as one GIVES CLARIFICATION. The same holds for REPEAT statements. Trivial clarifications are omitted; e.g., "O.K.?" "O.K.!" "You see?" "Yes."

CASE: A single case is scored as a single unit, no matter how many sentences it may take. The same rule holds for each elaboration of the case.

C. Context

In general, the context for scoring static operations is the statement scored. Most statements, then, could be scored if they were taken completely out of context.

Dynamic operations demand a context of at least two statements.

It should be noted that at least two categories in the static format actually involve dynamic operations. These are: CLARIFICATION and RELEVANCE.

CLARIFICATION, in particular, demands a larger context from which to score than the statement itself, except when there is an obvious cue.

Give credit for dynamic operations even when they are in response to someone else's statement or question.

D. Dynamic Operations -- Conventions

1. CONSISTENCY (c)

If the scorer can reasonably infer that a question asked will lead the antagonist into an inconsistent position, or lead him to explain an inconsistency, the question is scored CONSISTENCY. The frame of reference for ASK QUESTIONS, then, is whether the question would give rise to an inconsistency or would suggest an inconsistency in the mind of any average American.

However, the reply to such a question requires an explicit or near-explicit concern for the conflict or inconsistency in order to be scored CONSISTENCY.

In other words, if the intent of the questioner is to suggest an inconsistency, and he succeeds to the extent of persuading the scorer that an average person would see an inconsistency, the question is scored CONSISTENCY.

The maneuver the antagonist uses in avoiding or dealing with the inconsistency, however, is not scored as CONSISTENCY, unless it is obvious that the antagonist knows he is dealing with an inconsistency. He must say things like: "I know I am contradicting myself." "I know I said something different before." "This is the same position I have held all along."

If one antagonist suggests inconsistency through a CASE or rewording in a CLARIFICATION, and then goes on to press the other antagonist with "What decision would you make," or "Do you believe that?" both acts are scored as CONSISTENCY.

"Even though ..." is a highly probable cue to a c-score in a DIALECTICAL situation.

Posture three (QUESTIONS) is automatically scored c.

2. GENERALIZING and SPECIFYING (g)

- a. GENERALIZING involves making an empirical inference from more specific instances, but also includes drawing a conclusion from a number of facts.

The context of a GENERALIZING operation must always be at least two statements: one supporting or summarizing the other.

- b. The following static acts are valued mainly because they involve specification, and therefore the dynamic operation SPECIFYING is not scored:

SPECIFIC FACTUAL CLAIM OR STATISTIC
 SOURCE
 CLARIFICATION
 CASE

SPECIFYING generally occurs as one statement supporting another statement. It is important to identify more specific generalizations which are given to support more general claims, since the static category of GENERALIZING is given no value, while the dynamic act of SPECIFYING is.

Reasons given to support another position are scored GENERALIZING. --"I would desegregate the schools so that the Negroes would get a better education," is scored: SPECIFIC VALUE JUDGMENT + GENERALIZATION = SPECIFYING.

- c. GENERALIZING should not be scored unless it is clear that the speaker intends to operate at different levels of specificity.

4. QUALIFYING (q)

The basic idea behind qualifying is dealing with an exception which violates a general statement, i.e., modifying a statement to meet an exception. It is a low inference category; it should not be scored unless the individual is clearly aware that there is a conflict of values, claims, definitions, etc., and that he is dealing with this conflict.

Examples:

- VALUE JUDGMENT followed by "only under these circumstances would I do this."
 --LEGAL CLAIM: "Only under these circumstances will he not have a right to do this."
 --GENERALIZATION: "This holds true, except for these conditions."

QUALIFYING -- Conventions

- a. If several exceptions are considered in a single decision, or other category, this is scored as one statement.
 b. "Depends on" is a cue to qualification.
 c. Any statement scored q should also be scored c.
 d. If the word "qualify" (or a synonym) is used explicitly, this is scored as QUALIFICATION, whether or not the operation is performed

E. Static Categories -- Conventions

1. VALUE JUDGMENTS (GVJ:SVJ)

A VALUE JUDGMENT or decision is a statement which indicates that the speaker thinks that an object, person, action, or decision is good or bad, desirable or undesirable. Verbal cues are "should," "ought," "right" as a substantive adjective ("it is right"), etc. Decisions are also considered as value judgments, e.g., "We should do" Predictive statements about one's own behavior are value judgments: "I would do"

It should be noted that after the category CASE there is frequently ASKS FOR VALUE JUDGMENT followed by GIVES VALUE JUDGMENT.

GENERAL VALUE JUDGMENT (GVJ) vs. SPECIFIC VALUE JUDGMENT (SVJ): All value judgments or decisions are specific unless there is stated or described a more general value. General values include: personal freedom; the specific implications of personal freedom including property rights, contract rights, free speech, religious rights, and freedom of conscience; community, national, or general welfare; necessities of life including the right to work, the right to health, the right to eat, the right to shelter, the right to a decent living, etc. It should be noted that if the student says, "Someone has a right (or freedom)," this is scored as a LEGAL CLAIM. If, however, he says, "Someone should have a right to ...," this is scored as a VALUE JUDGMENT.

GENERAL VALUE JUDGMENT is scored only when it is clear that the student is appealing to some general ethical principle. If he says, for example, "Negroes should be treated well," this is scored as a SPECIFIC VALUE JUDGMENT. If he says, "Negroes should be treated equally," this is scored as a GENERAL VALUE JUDGMENT.

It is possible to paraphrase a general value. For example, if the student says, "Everyone should be treated on his own merits," the student would be given credit for a GENERAL VALUE STATEMENT because it is synonymous with the concept "equal treatment."

VALUE JUDGMENTS: Conventions

- a. "Nice" is a VALUE JUDGMENT; "beautiful," "ugly," "awful," "pretty," etc., are GENERALIZATIONS.
- b. When a student's general value position is rephrased or reloaded by the teacher or interviewer, score the teacher's statement CLARIFICATION.
- c. Statements of personal taste, feeling, or reaction are scored as VALUE JUDGMENTS.

- "I personally like Negroes."
 - "I would feel bad if I thought Negroes were never going to get a fair chance."
 - "I know it is wrong not to like everyone, but I still don't."
 - "I believe in"
- d. Distinction between GENERAL LEGAL CLAIMS and VALUE JUDGMENTS:
- "has a right to," "can," "freedom to," and the like followed by a general value are scored as GENERAL LEGAL CLAIM.
- "He has a right to freedom of speech."
 - "We have a right to freedom of press in this country."
 - "We can associate with whomever we please in this country."
- More specific rights are not scored as GENERAL LEGAL CLAIMS, but rather as SPECIFIC LEGAL CLAIMS. For example:
- "I could evict him if he didn't pay his rent."
 - "He can be required to bargain collectively."
- e. The statement "It is right to" is scored as a VALUE JUDGMENT (SPECIFIC or GENERAL, as the case may be).
- f. "I believe in" is scored as a VALUE JUDGMENT.
- "I believe in giving Negroes equal rights" is scored GENERAL VALUE JUDGMENT.
- g. "What do you say to that?" is scored as SPECIFIC VALUE JUDGMENT.
- h. "Deserve" is generally scored as a SPECIFIC VALUE JUDGMENT.
- It may be a causal GENERALIZATION, however, as "He sold them the house because they deserved it."
- i. "All right" is scored as a SPECIFIC VALUE JUDGMENT, except when used as a transitional phrase.
- j. "Would you mind if X (were to be done)?" is scored as ASKS FOR SPECIFIC VALUE JUDGMENT.
- k. "This is protecting the general welfare of all the people. Don't you think that is good?" Score two acts: first -- GIVES GENERAL VALUE JUDGMENT (because a general value was explicitly used -- otherwise scored as GENERALIZATION); second -- ASKS FOR GENERAL VALUE JUDGMENT (or SPECIFIC VALUE JUDGMENT, if general value is not mentioned).
- l. If a GENERAL VALUE JUDGMENT score has been given for use of a label (e.g., general welfare), subsequent references to the label are not scored GENERAL VALUE JUDGMENT unless the speaker is clearly committing himself to the value.

- "Labor unions defend the general welfare." = GENERAL VALUE JUDGMENT.
- "But they stand for the general welfare." = GENERALIZATION.
- "I think unions are good because they stand for the general welfare." = GENERAL VALUE JUDGMENT.

- m. "Can you think of any values involved in this case?" = ASKS FOR CLARIFICATION-ANALYSIS (a)
- n. "What do you say to that?": score SPECIFIC VALUE JUDGMENT
- o. When a more general value statement is derived from a specific instance (a judgment about all unions used to justify a judgment about a particular union, for example), and the student is still not at the level of a GENERAL VALUE JUDGMENT, score SPECIFIC VALUE JUDGMENT, g.
- p. "Does anyone have a plan?": score SPECIFIC VALUE JUDGMENT
- q. "I agree with Joe.": score SPECIFIC VALUE JUDGMENT
- r. Loaded language implies a value judgment.
- s. "Doesn't this seem reasonable?": score SPECIFIC VALUE JUDGMENT in DIALECTICAL context; score GENERALIZATION in DESCRIPTIVE context.
- t. "X is unnecessary.": score SPECIFIC VALUE JUDGMENT.

2. LEGAL CLAIMS (GLC:SLC):

LEGAL CLAIMS refer to real or hypothetical claims about the world which have a legal basis rather than a natural basis. The two most common language cues are "can" and "right to."

--For example: "You can drive only 45 miles an hour on Route 2."
 "He doesn't have any right to punch you in the nose."
 "If they suspend the Blue Laws, we will be able to keep the shop open on Sunday."

The distinction between GENERAL LEGAL CLAIMS and SPECIFIC LEGAL CLAIMS is essentially the same as that between these two kinds of VALUE JUDGMENTS. A GENERAL LEGAL CLAIM is a claim based on a more general legal principle. These principles have to be explicitly stated or clearly paraphrased. Examples of general legal principles are: Freedom of Speech and Religion, General Welfare, Equality, Property, Contract, Rule of Law, Constitutionalism, Due Process, Equal Protection of the Law, Separation of Power, Checks and Balances, Representation, Consent.

If the student stated, for example, that he has a right to buy a house if he has the money, this would be a SPECIFIC LEGAL CLAIM. If he said

this right is protected by the courts which would give him a fair hearing, he is clearly implying DUE PROCESS, and would be scored GENERAL LEGAL CLAIM.

If someone asks for the legal basis of a legal claim, this should be scored ASKS FOR GENERAL LEGAL CLAIM, since the student is implying that a more general legal principle is needed. Supporting a LEGAL CLAIM with reference to a specific law, however, would be scored as a SOURCE. If the student says, "Negroes have a right to vote because of the 15th Amendment" this would be scored GENERAL LEGAL CLAIM-SOURCE. It is scored GENERAL because "right to vote" is synonymous with "representation." If, on the other hand, the student said "Negroes can buy houses in white areas because they just passed the Fair Housing Practices Act," this would be scored as SPECIFIC LEGAL CLAIM-SOURCE. If the student said, "Negroes have their property rights protected under the Fair Housing Practices Act," this would be scored GENERAL LEGAL CLAIM-SOURCE.

"You can't break the law": scored as GENERAL LEGAL CLAIM.

"There is a law against that": scored as SPECIFIC LEGAL CLAIM.

3. Factual Claims (GEN:SFC)

Factual Claims are statements describing events, or generalizations based on the description of events. They include causal relationships, explanations, etc. A GENERAL FACTUAL CLAIM (GEN) involves obvious inference. A SPECIFIC FACTUAL CLAIM (SFC) describes an event that has a more-or-less specific beginning and end. A statistic is a SPECIFIC FACTUAL CLAIM. All information presented in the case which the student has just read may be considered as SPECIFIC CLAIMS, although it may involve inference.

WARNING: Do not confuse qualifications and causal generalizations. There is an important distinction between a causal GENERALIZATION and the dynamic operation QUALIFICATION. If the student says, "Negroes would get ahead if they were given a chance," this would be scored simply GENERAL CLAIM even though there is a conditional or qualifying phrase contained in it. The rule is that qualification will be scored only when the speaker deals with an exception to the GENERALIZATION. If the student stated, "All Negroes score lower on intelligence tests except those in the North," this would be scored as a GENERALIZATION-QUALIFICATION. This is different from saying, "If the Negro moves to Northern cities, he will score higher on intelligence tests." This statement is essentially a causal prediction. It is saying, under these conditions, this will happen. These statements are not scored as qualifications.

Factual Claims: Conventions

- a. Statements of personal taste, feeling, or reaction are scored as VALUE JUDGMENTS, even though they state facts about the speaker.

- b. "Can" is a factual claim if it refers to possibilities allowed by nature. ("I can see the wall." "We can all be taught to respect our neighbor.") "Can" is a legal claim if it refers to privileges conferred by law. ("You can drive 60 miles an hour on the Massachusetts Turnpike.")
- c. "Fault" is a factual claim. Beautiful, ugly, awful, adequate, etc., also denote factual claims. ("It was his fault.")
- d. All claims in the document read to student are treated as "specific factual claims" (SFC).
- e. References to the document read to student are treated as SPECIFIC FACTUAL CLAIMS. ("It said in the editorial")
- f. Personal decisions phrased in "would" form are VALUE JUDGMENTS. ("I would rent him the house." "I would have nothing against him.") Impersonal decisions phrased in "would" are general claims. ("They would" "People would")
- g. Causal claims preceded by a VALUE JUDGMENT are scored as two units. Causal claims preceded by a factual claim are scored as one unit.
 - "We should treat Negroes fairly because we brought them over from Africa." -- two units;
 - "Negroes don't do as well as whites in school because they have been deprived of so many things." -- one unit.
- h. Hypothetical questions are scored ASKS FOR CLARIFICATION, if the case upon which they are based has already been introduced. If a new situation is used, score as GIVES CASE + ASKS FOR SPECIFIC VALUE JUDGMENT.
- i. A contradictory claim is scored QUESTION CLAIM.
 - "Negroes are just as smart as whites." = GENERALIZATION
 - "But they score lower on I.Q. tests." = GENERALIZATION + SPECIFICATION + CONSISTENCY
- j. Inferences about how a law (or whatever) would operate if passed, or in a hypothetical situation, are scored GENERALIZATION.
- k. A challenge to a GENERALIZATION followed by a series of SPECIFIC FACTUAL CLAIMS is scored CHALLENGES SPECIFIC FACTUAL CLAIMS.
- l. Asking for evidence is scored as SPECIFIC FACTUAL CLAIM.
- m. "X is needed (necessary, etc.)" is scored as GENERALIZATION.

4. SOURCE

When a specific law or document is named, SOURCE credit is given for the first such reference only. SOURCE is not scored when the only

reference is by pronoun. ("It says")

"How do you know that?": scored as SOURCE

5. DEFINITIONS (DEF)

A DEFINITION is a statement about how a word or phrase is defined or should be defined. It is also a statement of analysis by which several meanings for a single word might be distinguished. For example, "The Soviet Union can be thought of as both state capitalism and socialism." Common language cues for definitional statements are "means" and "is."

- a. Restating in definitional form is DEFINITION rather than CLARIFICATION. For example:

--"Then you think that equality means that Negroes have a right to go to school with whites."

- b. Pure analytic statements which cannot be easily categorized are scored DEFINITION.
- c. When restating a position, score DEFINITION if the terms definition, define, etc. are used.

6. REPEAT

- a. Reference to a document under discussion and available to all is scored REPEAT.
- b. --In a DESCRIPTIVE context:
any highly structured question put by the teacher implying a definite answer is a cue to an answer to be scored REPEAT. For instance, if a given number of items, such as reasons, causes, etc., are asked for, score REPEAT. In general, the greater freedom the student has in answering and the more permissive the teacher is in asking, the less likely the response will be scored REPEAT.
--In a DIALECTICAL context, only "Yes" or "No" answers are scored REPEAT.
- c. "O. K.? Do you understand?": scored REPEAT.
- d. Any question about items the student is expected (by the teacher) to know (text, case, etc.), are scored REPEAT.
- e. When a student repeats an argument that has already been made, but is clearly thinking it out for himself, do not score REPEAT; score in appropriate categories.
- f. Do not score REPEAT with QUESTIONS (3); score in the original category of the statement being questioned.

- g. Agreeing, repeating, or leading the student on in a dialectical situation for purposes of trapping him in an inconsistency, score REPEAT-d.
- h. "That would be a different situation.": scored REPEAT-c.
- i. Statement of another person's position: score REPEAT.
 --"Ronny, you're the only person who disagrees."
 If Ronny questions this, score QUESTIONS REPEAT.
 --"Who says I disagree?"
- j. If an act is repeated, score REPEAT. Most instances when "Yes" is the response should be scored REPEAT.
 --Student: "I would rent the house to a Negro." = SPECIFIC VALUE
 JUDGMENT
 --Teacher: "You would rent the house to a Negro?" = ASKS FOR
 REPEAT
 --Student: "Yes." = REPEAT

7. CLARIFICATION (CLARIF.)

The general idea behind the category CLARIFICATION is an attempt by one person to communicate more precisely or more clearly to the other person something already under discussion.

CLARIFICATION includes: asking for explanation, reasons, elaboration, etc.; attempting to focus the argument on a more specific point; restating something said previously in different words. It should be noted that when one person rephrases something said by the other to change the loading of his original position, this is scored as CLARIFICATION. This holds also when the speaker summarizes his opponent's argument and in so doing, shifts the meaning and emphasis to change the value loading.

Analogies which are not dialectical in intention are scored CLARIFICATION.

Whenever there is doubt about whether to score CLARIFICATION or DEFINITION, score DEFINITION.

--Definition is an attempt to carry the argument along by making finer distinctions or pointing up differences between classes of events or cases.

CLARIFICATION -- Conventions

- a. A loaded clarification in which the student's words are phrased in such a way that they violate some commonly held value is always scored CLARIFICATION-CONSISTENCY.
- b. When the teacher infers from what the student has said that he is defending a more general position, and asks the student if this

is so, this is scored ASKS FOR CLARIFICATION.

--"Then you really think everyone should be treated equally?"

This convention holds only when one can fairly infer that this is a correct conclusion to draw from a number of previous statements.

--If the teacher begins a discussion with "Do you think everyone should be treated equally?" it would be scored as ASKS FOR GENERAL VALUE JUDGMENT.

--In both cases, the student's reply is REPEAT if the answer is a simple "Yes" or "No"; if more than that it is probably scored as GENERAL VALUE JUDGMENT.

- c. "This is a very complicated issue.": score CLARIFICATION-ANALYSIS (a).
- d. Correction of one's own statement is scored CLARIFICATION, unless strongly prompted by teacher (then score REPEAT).
- e. "Explain your position.": score CLARIFICATION-DIALECTICAL (d).
- f. When the course of the argument is being described abstractly (analytically), score CLARIFICATION-ANALYSIS (a).
--"Your last statement was analytical." "You really agree with me"
- g. This situation is just like the other one; they're both the same.": score GIVES CLARIFICATION.
- h. "That is," "In other words," are cues to CLARIFICATION-SPECIFYING.
- i. "Why?" "What reasons?" "How would you deal with?": score CLARIFICATION.
- j. "This just proves my point.": score CLARIFICATION-ANALYSIS (a).
- k. "We should qualify (specify, etc.) this position.": score CLARIFICATION + ANALYSIS (a) + g or g, etc.
- l. "Do you remember what we learned last year, that X is?": score CLARIFICATION.
- m. When a DIALECTICAL context becomes analytical, score the teacher's questions as CLARIFICATION-ANALYSIS (a); score student responses CLARIFICATION no analysis.
- n. "I didn't say X; I said Y.": two units -- score CLARIFICATION; REPEAT.
- o. "What do you say to that?" or any request for reasons: score CLARIFICATION.
- p. Score CLARIFICATION-ANALYSIS (a) for a suggestion as to the best way, of a number, to solve a problem.

8. CASE

CASES or analogies are used to define more precisely the speaker's or listener's use of a word or feelings about a decision or policy under discussion. They normally take the form, "Suppose this were the situation." They are often hypothetical, as indicated above; but they can be real. "For example, we had this situation in Concord. What would you have done about that, or how would you have felt about that?"

CASES are used primarily to clarify values and definitions; they always imply a DIALECTICAL posture. A person might say, "A democratic country is one which is ruled by the majority of all the people." This would be a definition. Someone might then say, "Suppose the majority of people in the United States decided to liquidate all the Negroes, as was done to the Jews in Nazi Germany?" This would be a case to clarify the meaning of the word "democracy" in a dialectical manner (i.e., it implies the existence of a certain value).

CASES are most often used to clarify the value position of the antagonists. A person might say, "I am against jailing Communists." The speaker might say, "What if valid evidence indicated that all Communists in America were involved in an active plot to overthrow the United States government by force?"

The category CASE is generally followed by a question in the category ASKS FOR SPECIFIC or GENERAL JUDGMENT.

The essential difference between a CASE and a CLARIFICATION is that the CASE is used to elicit from the other antagonist a decision, usually one that contradicts a previously held position.

CASE -- Conventions

- a. Replies to GIVES CASE take one common form: GIVES VALUE JUDGMENT, either SPECIFIC or GENERAL. ("I would do this." "People ought to do this.")
- b. Replies to GIVES CASE which take the form "That's different" are scored QUESTIONS CASE. The general rule is that they are scored QUESTIONS CASE when there is a vague reference to some difference.

If reply to GIVES CASE is "That situation has nothing to do with the discussion" this is scored QUESTIONS RELEVANCE.

If reply to GIVES CASE tells specifically how the second case is different from first, such as, "The right to buy a house is different from the right to buy drugs or groceries because . . .," it is scored GIVES DEFINITION, since an analytical procedure is taking place.

Elaborations or extensions of cases are scored CASE.

9. RELEVANCE

A statement which states, asks for, or questions whether a previous statement is relevant or central to the argument is scored RELEVANCE.

--"What does that have to do with the discussion?"
 --"We're not talking about X, we're talking about Y."
 --"This situation is like the other one in that both involve equal rights."

--"I don't see your point.": score QUESTIONS RELEVANCE.

--"That doesn't make any difference" is a highly probable cue for RELEVANCE.

--"Are these facts important?": Score ASKS FOR RELEVANCE.

10. DEBATE STRATEGY

Ad Hominem remarks which explicitly discuss the tactics being used by the opponent, rather than the problem under discussion are scored DEBATE STRATEGY.

--"You're trying to confuse me."
 --"You're just trying to get me off the point."
 --"You're pretending not to understand when you really do."

DEBATE STRATEGY -- Conventions

- a. "That's too bad" (sarcastically) and the like: score DEBATE STRATEGY.
- b. Flippant agreement to obvious immoralities (e.g., genocide): score DEBATE STRATEGY.

11. TASK PROBLEM: PROCEDURAL (TP:P)

A statement directed at controlling the immediate interpersonal situation, but said in good faith is scored TASK PROBLEM: PROCEDURAL. (That is, the statement does not assume that anyone is deliberately trying to foul up the discussion.)

--"Who wants to be chairman?" "Let's take a vote." "Let's give everyone a chance to talk."

TASK PROBLEM: PROCEDURAL-- Conventions

- a. When passing out or collecting material, the ensuing silence is scored TASK PROBLEM: PROCEDURAL, one unit every four seconds. All other forms of silence are scored as TENSION.
- b. ~~Personal~~ remarks or interpersonal exchanges such as "I'm tired"; "What did you do last night?"; etc. are scored TASK PROBLEM: PROCEDURAL.

12. DEVIANCE; DEVIANCE CONTROL

Score DEVIANCE whenever the student is clearly acting out of line. This should be fairly extreme; "wise guy" responses, for instance, are usually scored DEBATE STRATEGY.

A statement directed at controlling the immediate interpersonal situation, but assuming that the person to whom it is directed is violating the group norms, is scored DEVIANCE CONTROL.

--"Please be quiet and give someone else a chance to talk."
 --"You don't have to shout at me."
 --"Get back in your seat."

13. AFFECT

When affect is demonstrated, usually by tone of voice or use of positive or negative emotive language, this should be scored with one or two pluses or minuses. Some statements are purely affective -- e.g., "Very good, Eleanor." These should not be scored as separate units; score in box representing the nearest previous or subsequent act for the speaker.

Any indication that a previous statement is wrong, score single negative affect.

Use of student's name, except for initially calling on him, should be scored single positive affect, unless clearly negative.

14. TENSION/TENSION RELEASE

Silence is considered as a sign of tension; score TENSION with a check mark (✓) for every four seconds of silence.

Laughter is considered TENSION RELEASE (TR); score TENSION RELEASE with a check mark for every four seconds of laughter.

F. MISCELLANEOUS CONVENTIONS

"How do you know that?"

--If this follows a GENERALIZATION, score ASKS FOR SPECIFIC FACTUAL CLAIM.

--If this follows SPECIFIC FACTUAL CLAIM, score ASKS FOR SOURCE.

--If any other category, score MISCELLANEOUS.

Trivial rapid interchanges that act as acknowledgment of one another's statements, and which would ordinarily be scored as CLARIFICATION or REPEAT, are not scored.

Incomplete sentences are not scored.

"Right?" is scored as ASKS FOR _____ (whatever category is asked for).

"The Constitution says that all men are created equal.": score SOURCE + GENERALIZATION.

"The Constitution says that all men should be treated equally.": score SOURCE + GENERAL VALUE JUDGMENT.

--In each case the person has interpreted the Constitution rather than repeated it. A quote or paraphrase from a document is scored SOURCE + SPECIFIC FACTUAL CLAIM.

When several questions are asked, and then a reply is asked for, score ASKS FOR _____ (category of most recent question).

If category labels describing dynamic operations are used (e.g., qualify, specify, etc.), score as such whether or not the operation is actually performed.

When in doubt about scoring, score in more valued category or in less often used category.

In a conflict between two categories, score more valued category.

Any act that cannot be scored in another category, score in REPEAT.

Recognition by the teacher of a wrong or irrelevant student response is scored in the appropriate category with small negative affect.

In the following situations, Y refers to a statement made by the student (after a VALUE JUDGMENT), and X refers to a statement made by the teacher such that X is meant to demonstrate an implied contradiction in Y.

--"X is like Y," and the like, score ANALYSIS.

--"X is different from Y," score QUESTIONS in the category in which X was scored (e.g., QUESTIONS, CASE).

--"X means this," and any statements regarding the exclusion or inclusion of X in regard to category Y, score DEFINITION.

Idiosyncracies of speech which might be scorable are scored only if they seem to serve genuine cognitive functions. If a person says "I don't know" habitually before going on to make a statement, do not score SELF-DOUBT unless it seems definitely warranted.

TABLE C-1

NORMS DATA FOR TEACHER AND STUDENT RESPONSES AS CATEGORIZED BY A
CONTENT ANALYSIS SYSTEM FOR RECITATION AND SOCRATIC STYLES OF TEACHING

Response Categories	Teacher Behavior		Student Behavior	
	Rec.*	Soc.*	Rec.	Soc.
A. <u>Statement Types</u>				
General Value Judgment	.002	.019	.005	.021
General Legal Claim	.003	.010	.004	.018
Specific Factual Claim	.024	.040	.028	.023
Source	.010	.007	.006	.007
Definition	.023	.023	.016	.019
Clarification	.230	.173	.074	.061
Case	.006	.024	.002	.006
Relevance	.005	.008	.003	.004
Specific Value Judgment	.034	.134	.066	.204
Specific Legal Claim	.028	.034	.032	.062
Generalization	.078	.154	.214	.273
Repeat	.396	.239	.528	.281
Debate Strategy	.004	.005	.006	.023
Task: Procedural	.040	.024	.008	.010
Deviance Control	.006	.008	.001	.002
B. <u>Social-Emotional Responses</u>				
Affect: Positive	.124	.046	.000	.009
Affect: Negative	.044	.050	.004	--
Tensions**	.028	.046	--	--
Tension Release**	.021	.027	--	--
C. <u>Statement Posture</u>				
Questioning	.020	.023	.010	.014
D. <u>Logical Operations</u>				
Generalizing-Specifying	.005	.039	.026	.039
Consistency-Inconsistency	.032	.229	.010	.023
Qualification	.001	.002	.001	.005
E. <u>Discussion Posture</u>				
Dialectical	.046	.272	.003	.006
Analytical	.056	.033	.004	.006

*The recitation percentages are based on mean scores of a total of ten discussions for four different teachers; the socratic percentages are based on mean scores of a total of nine discussions for four different teachers.

**Tension and Tension release are classroom scores and were not divided here according to teacher or student.

CHAPTER THIRTEEN

STUDENT PERSONALITY AND THE ANALYSIS OF PUBLIC CONTROVERSY

The results of the experiment comparing socratic and recitation teaching reported in Chapter Twelve are not surprising. The large body of literature on methods research, especially that comparing student-centered with teacher-centered teaching, demonstrates that a single dimension of style does not have sufficient overall effect to produce a clear superiority of one style over another. Moreover, we had similar evidence from our own pilot experiment carried out in the spring of 1958, although the learning outcomes were not measured as carefully as in the major study. Although we did not anticipate that there would be substantial differences between the general effectiveness of the two styles of teaching, we did believe that different types of students would react to the two procedures in quite different ways, and that if it were possible to infer these differential reactions from quantifiable personality variables, we might find substantial differences in the results the two approaches had with particular students. While there is little research dealing with the interaction between student personality and teaching method, there is enough to suggest that it is a fruitful line of investigation. Wispe,¹ for example, carried out an experiment in which one group had a great deal of free discussion in section meetings connected with a course in elementary psychology, while the other group had highly structured section meetings. Neither method proved more effective in preparing students for a final objective examination. He did find, however, that students who valued personal independence were both critical of the structured sessions, and preferred the free discussions, while the converse relationship held for students who liked organization and structure. According to McKeachie, Patton² made a similar finding. Both experiments demonstrate, however, only that students like different types of classroom environments; not that different environments have an effect on learning. Glaser's classic study on critical thinking, while not varying teaching style, does offer some clues concerning which students can function best in a critical thinking atmosphere. Glaser administered a questionnaire to assess the "happiness" or emotional satisfactions of students who participated in a critical thinking experiment, and related the results to the gain in critical thinking ability. The questionnaire contained such items as:

- Do you regard yourself as religious?
- Do you feel appreciated by your family?
- Do you feel satisfactory adjustment to the opposite sex?

He found that those students who made the greatest gains in critical thinking consistently answered "no" to such questions in greater numbers than those who made smaller gains. Glaser comments:

¹L. G. Wispe, "Evaluating Section Teaching Methods in the Introductory Course," Journal of Educational Research XLV, 1951, pp. 161-186.

²J. A. Patton, "A Study of the Effects of Student Acceptance of Responsibility and Motivation on Course Behavior" (Unpublished doctoral dissertation, University of Michigan, 1955).

This finding may perhaps be explained on the ground that the pupils who gained the most in critical thinking scores were, as a group, also found to be intellectually superior to those who gained least. Persons who possess a combination of relatively high mentality and manifest a readiness to respond to training in critical thinking may well be expected to possess greater sensitivities, subtler and more varied personality needs, a sharper awareness of the nature of their satisfactions, and more resistance to saying yes than persons who possess a relatively low mental level and lack of readiness to respond to instruction in critical thinking. Those who gained most, then, probably are not as easily "satisfied" as those who gained least.³

Glaser's finding might be interpreted to mean that those students who tend to be more critical of themselves and their culture and/or are more sensitive and intelligent tend to respond to a "critical thinking" educational environment in a different way than those who are more conventional or less intelligent. This response, moreover, is translated into learning outcomes. This suggests that both intelligence and some attitudinal factor are related to responsiveness to a critical thinking environment.

Stern and his colleagues at Syracuse and Chicago report efforts to identify what attitudinal or personality factors play a key part in the student's perception of a critical thinking environment.⁴ Several studies were carried out using four personality types: the "authoritarian," the "antiauthoritarian," the "rational" and the "irrational." These typologies are made on the basis of a self-reporting inventory of beliefs, a modified version of the Inventory of Beliefs developed for the American Council on Education Cooperative Study of Evaluation in General Education.⁵ According to Stern:

Sixty items [of the revised Inventory of Beliefs] are nativistic-fundamentalistic generalizations that are parallel to but not identical with items of the California scales. The remaining forty items are reversals obtained by writing what might be called internationalistic-relativistic generalizations, rejected by the authoritarians but acceptable to antiauthoritarians. An example of an authoritarian item would be:

³Edward M. Glaser, An Experiment in the Development of Critical Thinking (New York: Bureau of Publications, Teachers College, Columbia University, 1941), p. 158.

⁴George A. Stern, "Environments for Learning," The American College, edited by Nevitt Sanford (New York: John Wiley and Sons, 1962).

⁵P. L. Dressel and L. B. Mayhew, General Education: Explorations in Evaluation (Washington, D. C.: American Council on Education, 1954).

"The many different kinds of children in school these days force teachers to make a lot of rules and regulations so that things will run smoothly."

A comparable antiauthoritarian reversal was:

"More playgrounds and fewer strict fathers would eliminate juvenile delinquency."⁶

Students with high scores on the authoritarian scale were classified as authoritarians. Students who dogmatically embrace items on the anti-authoritarian scale were classified as antiauthoritarians. Students who tended to accept either type of item were classified as irrationals, while students who rejected both were classified as rationals. (It is interesting to point out the similarity between Glaser's dissatisfied nonconformists and Stern's "rationals.") Using this classification Stern demonstrates that colleges have different proportions of these various types of students. Gladstein using the same typology, shows, according to Stern, that different types of students do in fact have different attitudes toward college and different study habits.⁷ Stern suggests that the various types of students were "evidently reared differently, and obviously have different viewpoints about a wide variety of things" He then raises a number of questions concerning whether or not authoritarians, antiauthoritarians, rationals and irrationals actually behave differently in the classroom, especially in humanities or social science classes where ideological issues might arise. Stern describes an experiment carried out by Stern and Cope which bears directly on this question:

Three special classes were set up in the citizenship course taken by all liberal arts freshmen at Syracuse University. One of these sections was composed exclusively of authoritarians, the second of antiauthoritarians, and the third of rationals, as determined by scores on the Inventory of Beliefs which had been administered before registration to all incoming freshmen. All three sections were taught by the same instructor who had agreed to a special assignment of students but was otherwise unaware of the particular characteristics of the three classes.

The instructor met with each of these groups once a week throughout the semester, and maintained a diary of the events taking place during these meetings. By the end of the first week he commented spontaneously on his recognition of the three groups of students as types that he had encountered before.⁸

⁶George A. Stern, "Environments for Learning," pp. 692-693.

⁷G. A. Gladstein, "The Relationship Between Study Behavior and Personality for Academically Successful Students" (Unpublished doctoral dissertation, University of Chicago, 1957).

⁸George A. Stern, "Environments for Learning," pp. 697-698.

The instructor's impressions of these three classes are what one would have predicted, and constitute a striking validation of the measure. The authoritarians are described as "interested in religion," "lacking in curiosity or initiative," and reticent about getting into class discussion. The anti-authoritarians are hypercritical, even going so far as to question procedures and readings used in the course. The rationals are described as polite, friendly; the obvious "future campus leaders." The instructor's general impression of the three groups remains very much the same over the whole semester, although a breakthrough with the authoritarians and the rationals is apparently achieved during the seventh week, when the instructor defends the institution of slavery. At this time both groups became sufficiently indignant to "fight back."

This experiment is especially important for at least two reasons. First, the explicit objective of the instructor was to teach the students to deal with ethical and political controversy. Second, the authoritarians who were isolated and therefore treated in a special way by the instructor did considerably better on an objective examination than their authoritarian counterparts who were assigned to sections randomly. Moreover, the superiority does not seem due to the quality of the particular instructor involved because the other two sections (rationals and anti-authoritarians) did no better than those in regular sections. Stern points out very clearly the essential significance of the experiment:

The significance of this study does not lie in the use of the discussion method as an educational panacea, but rather in the effect that the persistent application of particular discussion techniques had in helping a group of authoritarian students increase their knowledge in an area to which they are usually resistant.⁹

A productive line of attack on the analysis of the classroom would seem, then, to involve three major factors: the particular type of content which the teacher wants to handle, the discussion procedure used by the instructor, and finally the personality of the student. In the experiment described in Chapter Twelve instructors handled the same content with radically different styles and came up with no measurable overall difference between groups. This further supports the hypothesis that the interaction between teacher style (or discussion method) and student personality is of critical importance. Stern's work suggests that the ideological aspect of the belief system of the student is one important dimension of personality which must be taken into account.

McKeachie¹⁰ reports a series of experiments explicitly investigating the interaction between personality variables and teacher style variables as they

⁹Ibid., p. 701.

¹⁰W. J. McKeachie, "Motivation, Teaching Methods and College Learning," Nebraska Symposium on Motivation edited by Marshall Jones (Lincoln: University of Nebraska Press, 1961).

affect classroom achievement. Using the McClellan-Atkinson motivation theory as his major source of personality variables, he demonstrates a significant three way interaction among need for affiliation, warmth of the instructor as perceived by the student, and grades in an elementary psychology course. He demonstrates a similar relationship among achievement cues in the classroom, need for achievement, and grades. Other such complex relationships are discovered in other courses, but only for men. While the substance of McKeachie's work is less relevant to our own than that of Stern, the methodology is similar and important. McKeachie himself points out that:

One possible partial explanation for the meager findings in both these areas is that teaching methods affect different students differently. Students who profit from one method may do poorly in another, while other students may do poorly in the first method and well in the second. When we average them we find little overall difference between methods and no overall effect of a single motivational measure.¹¹

In our initial efforts to discover various factors which might be related to the student's ability to learn a critical analytic approach to controversial issues, we used a simple correlational approach. That is, we administered a series of measures in the general areas of ideological orientation, interpersonal orientation, and cognitive structure, and correlated the results with measures of critical thinking and what we called critical retention, i.e., the ability to retain without distortion facts relating to a controversial document. This approach proved fruitless for at least two reasons: many of the measures were inappropriate for a junior high school population; and the correlations between personality variables and critical learning, although significant, were so low that when the contaminating factor of intelligence was taken into account, consistent relationships became unstable and could not be interpreted.

At this point we decided that the only way to demonstrate any relationship between student personality and training in critical analysis, if indeed it existed, was to conduct experiments with a correlational-experimental design. This would take into account two additional considerations: 1) We would be able to use a learning outcome measure or growth index as a criterion measure, rather than some static estimate of "critical analysis." 2) We would be able to identify differential reactions to dramatically different instructional treatments by students with different types of personalities.

Two Pilot Experiments

In an effort to discover more complex relationships between teaching style, student personality, and learning outcomes two experiments were carried

¹¹Ibid., p. 112.

out, more or less independently, but with a similar problem. The first was carried out at an early stage of the project, and centered on the initial efforts to teach the unit on school desegregation. In this experiment we administered the following series of personality measures:

- a measure of toughmindedness
- a measure of tendermindedness
- a measure of need for affiliation (self-reporting)
- a measure of need for ascendance (self-reporting)
- Guilford-Zimmerman Temperament Survey
- the tendency to respond to nouns on the semantic differential in terms of descriptive extremes -- this was called a "dichotimization" measure.
- sociometric choices
- measures of need for aggression, affiliation, and recognition through thematic apperception test procedures
- attitude toward segregation

A number of learning outcome measures were administered before and after the teaching of the unit which lasted some three weeks. Four of these learning outcome measures are described briefly below:

a) Interest

Interest in the segregation-integration problem. A measure of interest was developed under the guise of a current event exercise. Students were asked to read six paragraphs abstracted from current news. After reading a news item, the student was instructed to write a one sentence precis. It was emphasized that the students could read the news items in any order they wished. Presumably equivalent forms of the test were administered eight times, three times before the unit was taught, twice during the teaching of the unit, and twice after the unit had been completed. The rank order of the summary of the integration-segregation story was called a measure of interest. In order to increase reliability the rank orders of the segregation-integration items from the three trials before the experiment were summed. This is the pre-test score. The trial immediately following the experiment was used as a post-test score. The last trial was called a post-post-test score.

b) Sensitivity to Evidence

This was a measure of the extent to which attitudinal items loaded with specific factual statements either for or against integration might influence the student's overall attitude toward the item. In half the items the evidence favored a pro-integration position; in the other half the evidence favored an anti-integration position. The score consisted of the difference between the rating of pro- and anti-integration items. The measure was given before and after the unit on school desegregation.

- c) Multiple Choice Test of Factual Information on school desegregation.
- d) An open-ended short item test of general conceptual information. This sometimes involved giving reasons, or writing brief essays. Scores were derived from systematic content analysis.

Eight variables were derived from a systematic observational instrument in which an observer categorized the student's non-verbal behavior into one of 16 different categories while class discussions were taking place. The number of behaviors observed in each category was expressed as a percentage of the total number of behaviors recorded for that particular subject. The arc sine transformation was applied to the percentages. Categories in which there were low frequencies were combined. The unit of observation was derived from a scanning technique, in which the twelve members of each group were scanned continuously, once each minute. The eight variables derived from the scheme are listed below and are self-explanatory:

passive withdrawn

attentive classroom behavior

raises hand

participates

speaks out (without being recognized)

emotional excitement over discussion

withdrawn with tension

tension release (fidgeting, doodling, laughing, etc.)

disruptive

The basic design of this experiment was essentially the same as the experiment described in Chapter Twelve, but the teaching treatment was only of two weeks' duration. We used three split groups, constituting six small groups. Three were taught with systematic socratic procedures; three were taught with systematic recitation procedures. Three different teachers were used.

Findings

As stated before, there were no overall differences in method. Socratic students did no better or no worse than recitation students on any of the learning outcome measures, although the difference between the two groups on the interest measure approached statistical significance in favor of the socratic style.¹²

¹²Socratic groups participated more actively and showed significantly less passive behavior than recitation students, but this was construed as a process variable, not a learning outcome variable.

The correlation between interest growth (pre-test and first post-test estimate) and toughmindedness for the recitation group was -0.23, while for the socratic group it was 0.46. This would lead us to believe that students with a toughminded approach to the desegregation issue gained significantly more in interest in a socratic situation than did the same type of students taught by a recitation method. A related finding comes out of the Guilford-Zimmerman scale labeled "general activity." The correlation between interest growth (as measured by the difference between pre-test and delayed post-test measure) and "general activity" is -0.36 for the recitation group and 0.36 for the socratic group. This is not surprising since we also found that "general activity" is quite highly correlated with Guilford-Zimmerman's "ascendance" (0.43) and "sociability" (0.57). One suspects that there is a general personality trait which might be characterized as toughminded extraversion, and that students with this kind of approach to social situations enjoy the socratic discussion, while less aggressive students are more favorable to the less threatening recitation procedure. It should also be noted that students who made the greatest gains in interest in the socratic groups are least prone to dichotomize, as measured by choosing adjective extremes on the semantic differential ($r = -.61$). This high negative relationship is not found in recitation groups ($r = .03$).

Differences were also found between treatment and personality variables on the sensitivity to evidence measure. There is a positive relationship between sensitivity to evidence (the student is more likely to switch his position on the desegregation issue in the light of specific loaded facts) and toughmindedness for the students taught in socratic groups (0.32), but a negative relationship for students taught in recitation groups (-0.22). This can be interpreted to mean that toughminded students tend to react quite differently to evidence which may challenge their position on the desegregation issue, depending on the kind of treatment to which they have been subjected. They become more sensitive in socratic groups and less sensitive (more dogmatic) in recitation groups.

Moreover, growth in sensitivity to evidence bears a striking positive relationship to change scores on the multiple choice fact post-test (0.47) for socratic students and a negative relationship on the same post-test (-0.31) for recitation students. Thus it would appear that the students who learned more facts in the socratic groups also tended to become more sensitive to the relationship between the loading of the fact and a position taken on the desegregation issue, while the reverse effect tended to take place in the recitation groups: Students who remembered more facts tended to take a more dogmatic position on the desegregation issue.

It should be clear that either of these two results (ignoring facts which tend to refute one's position, or shifting one's position in the light of new facts) can be considered "good" from a societal point of view. The more important question relates to the extent to which the person disregards the negative consequences of his decision after he has become aware of them, or the extent to which more subtle psychological defenses cause the individual to conveniently forget such negative consequences. It is this latter

interpretation we would tend to place on these results. It was obvious that students in the recitation groups did not take many of the facts which might be construed as the "negative consequences" of desegregation seriously. (Facts such as the crime rate, disease rate, and illegitimate birth rate among Negroes.) They simply did not believe them. The students subject to socratic teaching, however, were required to deal with these stubborn facts in a controversial setting. It is quite likely that they came to believe them after they had been taught in this setting, while the recitation students simply ignored them or "forgot" them. In general, however, it should be remembered that recitation students did as well on the fact tests as did the socratic students. The significant finding is that those students who learned the facts best in socratic treatment tended to be the same students who were least stable in their position on the desegregation issue. And the reverse was true in the recitation groups. It is not that students were not aware of the facts which applied to the decision on the issue when they took the "sensitivity to evidence" measure. The facts are incorporated into the items on the test. The most reasonable interpretation would be that the students who learned the most facts in the socratic discussions came to believe these facts, while the students who learned the most facts in the recitation discussion treated them as outside their framework of personal decision.

These findings indicated that it was plausible to assume that the greatest research payoff would come from a research design which compared not only one teaching procedure or style with another, but took into account personality differences of individual students. The two major deficiencies in the experiment, however, were the brevity and the weakness of the dependent measures. One might get measurable effects over the course of three weeks, but in a longer period of teaching would different kinds of students accommodate themselves to different teaching styles, or would the differences be exaggerated? Moreover, the findings were based primarily on measures of interest, sensitivity to evidence, and information retention. None of these measures expressed well the more complex type of intellectual processes we had as our long-range objective.

Resnick's Study

Resnick¹³ conducted a much briefer experiment than that described above, but with a similar objective: to see whether or not certain personality measures would predict accurately differential student reactions to Socratic stress. The personality measures under consideration were in the area of cognitive style (openmindedness, the use of broad categories rather than narrow categories, communication among analytic categories, etc.). As a result of a factor analysis of a large battery of measures to get at these two cognitive traits, she came out with four factors:

¹³Lauren B. Resnick, "The Handling of Contradictory Information: A Study in the Effects of Cognitive Need, Cognitive Style, and Motivational Arousal," 1962 (Mimeographed).

Factor I -- A desire to do well in school

Factor II -- Size of categories used to structure messages

Factor III -- Femininity

Factor IV -- Production of Transformations

An experimental procedure was then set up to test whether students with different factor scores would react differently to a Socratic treatment as compared to those who had no such treatment. Specifically, the students were told that the experimenter would be interested in learning about their opinions on a currently important social question, socialized medicine. To help them make up their minds, they were given a paper to read describing the issue in an objective fashion. The paper was carefully constructed with an equal number of arguments on both sides of the issue. Furthermore, it was stressed that there was no right side to the issue. Any point of view was acceptable. In the control groups (45 students in two school classes), the paper on socialized medicine was distributed, and the students given the remainder of the class period (20 minutes) to read and formulate their thoughts. In the experimental groups (45 students in two high school classes) a similar procedure was carried out. In addition, however, the teacher asked for the students' oral opinions on the subject. After these opinions were solicited, there followed a brief Socratic discussion based on two dilemmas: (1) Are all men of equal worth, or is the higher status member of a community more "important"; and (2) Is the man who is likely to pay later entitled to better care than the man who is a "charity" patient? After about 15 minutes of Socratic questioning, the paper on socialized medicine was given to the students, and they were then given 20 minutes to read it. Three learning outcome instruments were administered the following day to get at a number of variables, among them the student's position on socialized medicine, openness (the ability to see both sides of the question and suspend judgment), opposition (the extent to which the student treated contradictory statements as a unit), slant (the extent to which the student considered statements which were affectively loaded in only one direction), and percentage of statements in the initial message contradicting the student's point of view that the student chose to answer.

While a number of interesting findings came out of this study, one of the most startling was that the relationship between Factor IV (production of transformations) and several learning outcome variables was quite different for the control group than for the experimental group. Resnick comments on these results as follows:

....it is clear that under neutral conditions, the predicted relationships held; that is, the Production of Transformations was associated with "open-mindedness." Finally, it is apparent that the Experimental arousal treatment caused marked differences in the effect of the Production of Transformations factor on cognitive behavior. In effect, the Socratic discussions seems to have made subjects who were able to produce transformations

TABLE 13.1⁴

**CORRELATIONS BETWEEN STUDENT'S SCORE ON PRODUCTION OF
TRANSFORMATIONS TEST AND FOUR LEARNING OUTCOME MEASURES FOR
SOCRATIC AND CONTROL GROUP**

	<u>Learning Outcome Measures</u>			
	<u>Openness</u>	<u>Opposition</u>	<u>Slant</u>	<u>Percent Statements Answered</u>
Socratic Arousal Group	-0.31	-0.43	0.17	-0.30
Control Group	0.41	0.42	-0.32	0.22

*This table is abstracted from Table 9 in Lauren Resnick, "The Handling of Contradictory Information: A Study in the Effects of Cognitive Need, Cognitive Style, and Motivational Arousal."

behave in "close-minded" ways in regard to the issue of socialized medicine. It would appear that certain kinds of motivational arousal, at least, can cause generally "open-minded" individuals to turn "closed-minded" with regard to the issue at hand.¹⁴

Resnick then goes on to suggest the following interpretation:

Perhaps the Socratic discussion leads subjects with the capacity to make transformations to quickly explore a number of possible "solutions" to the problem raised. Since, both implicitly and explicitly, the Socratic procedure demonstrates the inadequacy of these solutions, subjects who have considered them, may come to the conclusion that there is "no way out." As a result, when the time comes to express their opinions on paper, they may, in effect "give up" on the problem, and simply support one side or another without further attempts to consider opposing points of view. Subjects who make transformations less quickly, on the other hand, and have not yet explored all the avenues, still find the problem a real one and deal with it in a more serious manner in written form.¹⁵

Resnick's study is particularly significant in that while it set out to explore both motivational and intellectual variables associated with a "critical" or "openminded" view toward a social issue, it did not find an interaction between motivation and production of transformation, but rather an interaction between type of motivational arousal, the student's ability to produce transformations, and certain measures related to "open-mindedness." This, in a sense, provides us with a link between the previous studies, which concentrated mainly on social-emotional personality variables, and a study which was concerned mainly with an intellectual function. Thus Glaser found that high school students who were more dissatisfied with their surroundings were more intelligent and made greater gains as a result of a "critical thinking" treatment. Stern and Cope found that anti-authoritarians were more responsive to a Socratic type of teaching, and that only after sustained treatment did the "rationals" and "authoritarians" become responsive. (Anti-authoritarians also tend to be more intelligent.) In our own pilot study we found that students with a more "toughminded" attitude toward how the desegregation problem should be handled were affected more by Socratic treatment in that they developed a more relativistic view toward the problem. Resnick now finds that the more intelligent student, who is especially good at seeing complex verbal relationships tends to freeze after an immediate one shot Socratic arousal. This finding makes sense, however, if one distinguishes between the limited one shot "arousal" situation and the longer term educational treatment.

¹⁴Ibid., p. 25.

¹⁵Ibid., p. 27.

It may well be that the person who is more sensitive to abstract relationships sees more quickly the complexity of the situation with which he is asked to deal, and therefore, when faced with real value dilemmas which he has never been taught to handle, simply withdraws. On the other hand, it may be precisely this same type of person who is most teachable over a longer period of time.

Personality Measures Administered in the Concord Project

In the light of the work described above by Stern, McKeachie, Resnick, and others it seemed worthwhile to explore the relationship between selected personality factors and the student's ability to understand and apply the analytic and legal-ethical concepts described in Part Two. Broadly speaking, we selected three types of personality measures: Measures of cognitive need and cognitive flexibility; measures of ideological commitment; and comprehensive measures of social-emotional traits. These measures are described briefly below:

Cognitive Need and Flexibility

1. Resnick Self-Reporting Need-cognition Questionnaire

Sample Items:

If some music comes on the radio that you like, but don't recognize exactly, are you likely to just sit back and enjoy it or to try and figure out what it is, who might have written it, or who is performing?

Suppose you had a choice of spending a week alone in the country with a friend of your own sex. You have a choice of two people. One is a "nice guy" (or girl), easy to talk to and easy to be with. He is a lot of fun, but doesn't have very many ideas of his own. He isn't willing to talk seriously for very long. The other person is somewhat difficult to get along with. He is moody and touchy. But he has many interesting ideas which he likes to talk about and you find him very stimulating to talk with. Which person would you choose?

2. Wesley Rigidity Measure

Sample Items:

When I do math homework, I check each problem at least twice.

When I come across a word I don't know, I look it up in the dictionary.

3. Berlak school work-habit questionnaire

Sample Items:

I take notes when I read in school.

I write down my homework assignments in all my subjects.

4. Anderson Self-Reporting Need-Achievement Questionnaire

This is an attempt to assess the student's concern about doing well in school and succeeding in general.

Sample Items:

I would be very unhappy if I was not successful at something I had seriously started to do.

I always try to do at least a little better than what is expected of me.

Measures 1-4 were scored on a seven-point Likert type scale.

5. Semantic Differential Dichotomization Test

This measure was patterned after Osgood's Semantic Differential in format, and the variable was obtained by counting the number of times the student chose to describe objects in extreme terms (choice of "1" and "6" responses). Example:

	<u>Rope</u>						
weak	1	2	3	4	5	6	strong
angular	1	2	3	4	5	6	rounded
rough	1	2	3	4	5	6	smooth

Half of the terms to be described were neutral and half were politically loaded, e.g., Communist. Our findings demonstrated that whether or not the terms were socially loaded made little difference in the student's responses. If he tended to dichotomize, he did this with both types of words. ($r = .76$).

Social Attitudes and Ideology

1. "Have" -- "Have not" questionnaire.

An effort was made to assess to what extent students were sympathetic toward economically underprivileged groups, and to what extent they thought that existing privileged groups should maintain their present status.

Sample Items:

People who own a house or business are usually more responsible citizens.

Generally, people are poor because they're too lazy to earn a decent living.

2. F-Scale

This is presumably a measure of the tendency toward authoritarian solutions to basic socialization problems and interpersonal relations.

Sample Items:

It is only right for a person to feel that his country or religion is better than any other.

No sane, normal, decent person could ever think of hurting a close friend or relative.

3. Opinion of Southerners

This measure was included as a general measure of distrust of an "outgroup."

Sample Items:

There is nothing about Southerners that I could ever like.

I don't trust Southerners.

4. Opinion of Jews

A measure of ethnic prejudice where, presumably, there has been considerable outside training. There were almost no Jews in the community in which the experimental work was done.

Sample Items:

I suppose the Jews are all right, but I've never liked them.

I respect and admire the Jews.

5. Negro Social Distance Scale

A standard social distance measure.

Sample Items:

I would not mind playing on a team in which there were Negroes.

I would not mind Negroes living in my city.

These five measures were again scored on a seven-point scale.

Comprehensive Personality Inventories

Two comprehensive personality inventories were administered: The Guilford-Zimmerman Temperament Survey and the Cattell High School Personality Questionnaire.

The traits presumably measured by these inventories are listed and defined briefly below.

Guilford-Zimmerman Temperament Survey

1. G -- General Activity: a high score suggests a rapid pace of activities, high energy, liking for speed, quickness for action.
2. R -- Restraint: a high score suggests serious-mindedness, persistent effort, and self-control.
3. A -- Ascendancy: a high score suggests leadership habits, speaking with individuals, non-submissiveness.
4. S -- Sociability: a high score suggests that the respondent has many friends, likes entering into conversations, likes social activities.
5. E -- Emotional stability: a high score suggests evenness of moods, optimism, good composure.
6. O -- Objectivity: a high score suggests that the respondent is thick-skinned: that he is not hypersensitive, self-centered, or oversuspicious.
7. F -- Friendliness: a high score suggests a toleration for hostile action and a tendency not to be belligerent, hostile, or dominating.
8. T -- Thoughtfulness: a high score suggests reflectiveness, meditativeness, and a philosophical inclination.
9. P -- Personal Relations: a high score suggests tolerance of people and faith in social institutions.
10. M -- Masculinity: a high score suggests interest in masculine activities and vocations, resistance to fear, and a tendency to be "hardboiled."

High School Personality Questionnaire

The following brief description of the HSPQ is taken from Table I of the Manual. Trait designation, as with the Guilford-Zimmerman Survey, is by letter.

1. A -- Stiff, critical, aloof versus warm and sociable
2. B -- Low general intelligence versus high general intelligence
3. C -- Emotional, immature, unstable versus mature and calm
4. D -- Phlegmatic and stodgy versus excitable and unrestrained
5. E -- Submissive versus dominant
6. F -- Sober and serious versus enthusiastic and happy-go-lucky
7. G -- Casual and undependable versus conscientious and persistent
8. H -- Shy and threat-sensitive versus adventurous and "thickskinned"
9. I -- Tough and realistic versus esthetically sensitive
10. J -- Liking group action versus fastidiously individualistic
11. O -- Confident versus insecure
12. Q2 -- Group dependent versus individually resourceful
13. Q3 -- Uncontrolled and lax versus controlled, showing will power
14. Q4 -- Relaxed and composed versus tense and excitable

Sociometric Status

In addition to the above measures, the students were asked six questions (three positively oriented and three negatively oriented) concerning judgments about their classmates. (In summary form) the three questions were:

1. Who would you most (or least) like to work with on a committee?
2. What person in this room would you most (or least) like to have as chairman of the committee?
3. What person in this room would be the most (or least) help in the work of such a committee?

In previous studies we have found that the intercorrelations among statements on this type of measure are so high, we used a single index of sociometric status.

Reliability of the Measures

Table 13.2 presents the reliability of the personality measures used. Since most of these estimates were obtained by correlating split-halves of

TABLE 13.2

RELIABILITY ESTIMATES FOR PERSONALITY VARIABLES*

Cognitive Need and Flexibility

1. Resnick need-cognition	.50
2. Wesley Rigidity	.43
3. Berlak School Work Habit	.70
4. Anderson Need-achievement	.46
5. S-D Dichotimization	.70

Social Attitudes and Ideology

1. "Have--Have-Not" Questionnaire	.41
2. F-Scale	.46
3. Opinion of Southerners	.15
4. Opinion of Jews	.36
5. Negro Social Distance Scale	.49

Personality InventoriesGuilford-Zimmerman

General Activity	.55
Restraint	.51
Ascendancy	.68
Sociability	.74
Emotional Stability	.68
Objectivity	.71
Friendliness	.71
Thoughtfulness	.65
Personal Relations	.21
Masculinity	.83

HSPQ

A Aloof vs. Warm	.24
B Intelligence	.39
C Emotional maturity	.33
D Stodgy vs. excitable	.36
E Dominance	.33
F Serious vs. happy-go-lucky	.37
G Casual vs. Conscientiousness	.32
H Shy vs. adventurous	.38
I Tough vs. sensitive	.45
J Groupiness vs. individualistic	.17
Q1 Confident vs. insecure	.45
Q2 Dependent vs. resourceful	.09
Q3 Lax vs. controlled	.22
Q4 Relaxed vs. tense	.00

*All reliability figures are uncorrected split-half estimates, except HSPQ which has two forms.

tests, they would be somewhat higher if the Spearman-Brown formula were applied. The most striking aspect of the reliability correlations is the low reliability of the HSPQ, especially compared to the G-Z Survey, which is essentially the same kind of measure. Comparing traits which have similar nominal constructs, we find, for example, that "Ascendancy" on the G-Z has a reliability of .68, while "Dominance" on the HSPQ has a reliability of only .33; likewise, "emotional maturity" or "emotional stability" has a reliability of .68 on G-Z and a reliability of only .33 on the HSPQ.

The reliability results on the social attitudes scales are actually surprisingly high, if anything. The only exception is the "opinion of Southerness" measure. Undoubtedly, there are two factors at work in the discrepancy between this and the other measures: familiarity with the attitude object, and the specificity of the questions. Junior high school students, for example, do not have a very specific concept of the term "Southerner." It should be noted, moreover, that when the format of the questionnaire shifts from general opinionnaire to social distance questions, the reliability tends to rise.

Relating Personality Data to Learning Outcomes

With some 35 personality variables to relate to 12 learning outcome measures, the possibilities for different kinds of analysis are, of course, enormous.¹⁶ We decided to attempt in our preliminary analysis a simple and straightforward quest for relationships. This resulted in a correlation matrix relating all personality variables to all learning outcome variables. The number of large or consistent relationships was surprisingly few, and those that did exist often seemed contaminated with the more general factor of intelligence. From the total matrix of 35 variables, "the fourteen most promising variables were selected for additional analysis. We decided that the first step would be to partial out what might be called "general test-taking ability." This was defined as an estimate of I.Q. and three tests from the Differential Aptitude Test Battery: verbal ability, abstract reasoning ability, and clerical skill. The results of this analysis are reported below in Table 13.3.

Table 13.3 reports three types of findings: relationships among personality variables; relationships among learning outcome variables; and relationships between personality and learning outcome variables.

Relationships among personality variables. Inspecting the correlations among the personality variables reveals two quite distinct clusters of variables. The first three variables, need-achievement, need-cognition, and the Berlak School Work Habits show significant and substantial intercorrelations. This is consistent with Resnick's finding, and suggests that we are tapping

¹⁶We expect to report many of our findings about the relationship of personality measures to one another in future papers. This, however, was beyond the scope of the Project.

TABLE 13.3 (Key)

1.	Feed-Achievement
2.	Need-Cognition
3.	Berlak school work habits
4.	HSFQ G (persistence)
5.	S-D Dichotomization
6.	G-Z E (emotional stability)
7.	G-Z O ("thickskinned")
8.	G-Z F (friendliness)
9.	G-Z A (ascendance)
10.	G-Z T (reflectiveness)
11.	G-Z P (personal relations and tolerance)
12.	G-Z R (restraint and serious-mindedness)
13.	Sociometric Status
14.	Opinion of Jews
15.	American History
16.	Principles of American Government
17.	ITED No. 1 (social studies concepts)
18.	ITED No. 5 (interpretation)
19.	WAGNIS (critical thinking)
20.	Headlines H (interest in topics studied)
21.	Headlines G (interest in topics not studied)
22.	SIAT No. 1 (boss)
23.	SIAT No. 2 (siat-life)
24.	SIAT No. 3 (non-soc interview)
25.	SIAT No. 4 (soc interview)
26.	Unit Tests

TABLE 13.3

CORRELATIONS BETWEEN SELECTED PERSONALITY VARIABLES AND LEARNING OUTCOMES WITH I.Q.,
VERBAL ABILITY, ABSTRACT REASONING ABILITY, AND CLERICAL SKILL PERTAINED OUT

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	
29	57	61	63	63	63	63	63	63	63	63	63	63	63	63	63	63	63	63	63	63	63	63	63	63	63	
39	61	63	63	63	63	63	63	63	63	63	63	63	63	63	63	63	63	63	63	63	63	63	63	63	63	
03	08	00	12	15	26	06	15	25	63	21	55	30	19	26	29	01	06	05	02	17	30	29	35	51	62	68
00	22	15	26	06	15	25	63	21	55	30	19	26	29	01	06	05	02	17	30	29	35	51	62	68	70	72
10	14	08	15	25	63	21	55	30	19	26	29	01	06	05	02	17	30	29	35	51	62	68	70	72	74	76
21	11	06	15	25	63	21	55	30	19	26	29	01	06	05	02	17	30	29	35	51	62	68	70	72	74	76
32	43	15	19	06	30	19	26	29	01	06	05	02	17	30	29	35	51	62	68	70	72	74	76	78	80	82
42	36	30	19	18	23	17	15	29	01	06	05	02	17	30	29	35	51	62	68	70	72	74	76	78	80	82
05	24	25	22	04	37	58	49	13	37	21	07	07	07	07	07	07	07	07	07	07	07	07	07	07	07	07
16	32	33	31	01	09	15	20	13	37	21	07	07	07	07	07	07	07	07	07	07	07	07	07	07	07	07
27	15	21	16	23	16	21	08	09	08	07	12	09	01	06	05	02	17	30	29	35	51	62	68	70	72	74
20	07	27	20	04	04	07	09	08	08	07	12	09	01	06	05	02	17	30	29	35	51	62	68	70	72	74
21	16	03	26	02	21	03	17	01	06	05	02	17	30	29	35	51	62	68	70	72	74	76	78	80	82	84
21	23	05	20	00	05	02	02	08	08	05	01	04	33	11	62	68	70	72	74	76	78	80	82	84	86	88
20	19	04	12	05	12	02	17	15	15	01	04	06	14	20	70	54	60	60	60	60	60	60	60	60	60	60
21	17	09	23	06	16	02	08	08	14	07	06	12	11	19	52	36	39	39	39	39	39	39	39	39	39	39
06	05	01	17	01	10	04	08	04	00	05	12	11	19	52	36	39	39	39	39	39	39	39	39	39	39	39
08	35	35	41	09	33	33	26	20	06	25	23	10	09	27	01	08	15	07	00	64	07	12	47	28	38	06
10	20	16	39	16	34	26	15	11	10	20	09	16	03	33	10	08	15	07	00	64	07	12	47	28	38	06
19	13	19	21	22	18	21	22	04	02	24	12	40	17	34	16	36	36	36	36	36	36	36	36	36	36	36
29	26	15	36	06	18	00	15	14	04	03	03	15	13	21	29	30	36	33	26	08	07	42	28	38	06	06
29	39	30	30	04	03	04	02	30	21	08	12	21	24	28	49	36	33	13	13	06	06	13	06	38	06	06
16	13	10	28	03	09	07	03	29	21	14	18	12	07	05	15	11	13	13	13	06	06	13	06	38	06	06
26	18	12	08	18	08	04	15	21	21	02	07	28	21	41	47	15	41	30	08	09	39	27	35	06	06	06

decimal points omitted

Factor I in her factor analysis. It is interesting, moreover, that a number of social-emotional measures correlate with this trait system: G-2 Ascendancy, Thoughtfulness or Reflectivity, and Restraint or Serious-mindedness. There is nothing surprising about the fact that these variables tend to cluster together. We get the picture of a student who is highly motivated to do well in school work and who is interested in learning in general. He tends to be systematic in what he does. He is also somewhat aggressive, thoughtful, and serious-minded. The only surprising fact is that the two traits of self-discipline and need-cognition seem to go together. Some educators and psychologists have speculated that the more creative individual (who might be assumed to have a high need-to-know) tends to be punished by the rigid institutional framework within which the school operates. Our data indicate that this tends not to be the case.

The second cluster of personality variables consists of G-2 Emotional Stability, Objectivity (being thick-skinned), Friendliness, and Personal Relations (tolerance of other people and ideas). The major construct underlying these variables is apparently an ability to get along with other people without feeling personally threatened.

It might also be noted that more persistent, better organized, and more reflective students tend to get a higher sociometric rating. This should be no surprise, since two of the three sociometric questions concern the ability of the student to carry out assigned work in a group.

The dichotomization measure shows a non-significant but consistent trend to correlate negatively with both of the clusters identified above. This is also, to some degree, true of the measure of anti-Semitism, which is significantly related to reflectiveness. There is, however, a surprising positive relationship between anti-Semitism and the Berlak measure (high need for structure in school work). This might easily have been predicted before the study, because of the frequently cited relationship between ethnic prejudice and rigidity. This does not show up in our work, however, and we tend to find instead a trend in which need for structure becomes a positive rather than a negative quality in the syndrome of the reflective student.

Relationship of personality variables to learning outcomes. Having discovered certain consistent clusters of personality variables described above, the question now becomes: To what extent do these traits or tendencies facilitate or inhibit the intellectual functioning of the individual so that he will respond better or worse to instruction in the analysis of controversial public issues? As the matrix in Table 13.3 indicates, the answer to this question is complicated.

It would appear that the measures of interest in social problems are affected by both school interest and motivation as well as emotional maturity. Both measures of interest correlate highest with the HNFQ factor G, which includes both perseverance and orderliness in the n-cognition and Berlak measures and some items related to emotional maturity. This is further evidence that both of these factors are related to interest in social problems. There

is, moreover, evidence that these two factors are related to intellectual achievement, especially as this is reflected in the SIAT measures Nos. 2 and 3. We would interpret this to mean that there is probably no direct relationship between emotional maturity and orderliness in school and the ability to do well on these tests. It is more probable that these factors intervene to facilitate learning. They are probably both general measures of "school adjustment." This may be especially true of the first cluster of measures (N-Achievement, Need-Cognition, Berlak, and HSPA G), since they tend to bear a positive relationship with almost all measures of achievement. The highest relationships happen to be the ones in which there was the most direct training.

A second finding is the rather consistent pattern of positive relationships between sociometric status and learning. This is not surprising, since we would suspect that sociometric status would be another index of school adjustment, and, like the two factors discussed above, bear a positive relationship to learning.

One of the most surprising and exciting findings is the significant positive relationship between measures of persistence (HSPQ G), ascendance (G-Z A) and the SIAT No. 4. It was noted above that No. 4 is related only to No. 3 and not to the two other SIAT measures. The question then becomes: What do SIAT No. 3 and SIAT No. 4 have in common which is not common to any of the other measures? Two answers immediately come to mind. Neither is as structured as the first two tests (the student is rewarded for spontaneous responses). And second, both are administered by personal interviews. We might speculate that doing well on SIAT No. 4 (the sociometric interview) may well depend not only upon one's training, but also upon the student's tendency to be outgoing and persistent. This, of course, raises all kinds of problems for the development of evaluation instruments. We have stressed throughout our work the necessity of measuring analytic skills in "natural" and less-structured settings. This, almost of necessity, means the use of oral behavior rather than written or precoded information. But if oral behavior is significantly affected by important personality factors, which may have little to do with competence in analytic thinking, we must find either a methodological or statistical means of partialing out these effects.

Some comments on the interrelationships among the learning measures. Among the most interesting findings revealed by Table 13.3 are the high correlations among the measures even with the intellectual ability factor partialled out. The correlation between the American History test, which demands specific factual memorization, and ITED No. 5, which demands sophisticated reading of complex materials, is .70. These data suggest that there are at least three components of the type of analytic thinking we were attempting to measure. First, there is general intelligence which, to some degree, is partialled out here. Second, there is a general verbal reasoning factor, which is separate from whatever is measured by general intelligence. And finally, there are the specific intellectual skills involved in the analysis of political issues. Some of the relationship between the four SIAT measures must get at this last factor, since all are sensitive to special instructional treatment.

As indicated above, however, unraveling the various components which add up to a "score" on one of these tests is a major task that still stands before us. Our own feeling is that present instruments are too crude to make additional analysis of existing data very useful. A more fruitful approach would involve the systematic development of a wide variety of types of experimental instruments and conditions of administration, and then proceed to test the instruments out on various validation groups, that is, groups about whom we have a considerable body of other evidence indicating a specific level of political sophistication.

Interaction Between Student Personality and Teaching Style

In the first part of this chapter we reviewed a number of studies which were concerned with the relationship between personality and learning outcomes. As a result of this brief review as well as the discussion of teaching style in Chapter Twelve, it became clear that traditional efforts to seek relationships between teaching style and learning outcomes and student personality and learning outcomes would probably be less successful than a more complex design in which both antecedent factors were taken into account. In an effort to test hypotheses regarding the interaction between student personality and teacher style we selected the 13 personality measures which seemed most clearly related to learning outcomes, in general, and did two-way analyses of variance with covariance adjustment for I.Q. and pre-test measures for seven learning outcomes. This yielded 91 two-way analysis of variance results. The measures used in this analysis are summarized in Table 13.4. All of these measures have been discussed in previous sections of Part Three.

Of the 91 relationships investigated, some 12 turned up with significant interactions at the 5 per cent level. With this number of runs, we would expect about five significant interactions simply by chance. The only reasonable approach to an interpretation of such a fishing trip, we feel, is to look for some consistencies across both personality and learning outcome measures. There appeared to be very little consistency for the measures of interest (Headlines G and H), so these findings were discounted as chance occurrences. Six of the thirteen personality measures bore sufficiently interesting results to warrant interpretation here.

Selected results from two-way analysis of variance. Tables 13.5 through 13.11 summarize the interpretable results of the two-way analysis of variance. Table 13.5 indicates a clear and consistent relationship between need-cognition and four of the five learning outcomes. This finding simply repeats that noted in the correlation matrix presented in Table 13.3. One interesting additional finding comes out in these data, however. This is the different pattern of achievement for socratic trained high need-cognition students on the unit tests versus the SIAT No. 4 (socratic interview) test. High need-cognition students trained by a socratic procedure do relatively less well on the unit tests than do high need-cognition students trained with recitation

TABLE 13.4

SUMMARY OF PERSONALITY MEASURES AND LEARNING OUTCOME MEASURES INCLUDED IN THE TWO-WAY ANALYSIS OF VARIANCE INVESTIGATION OF THE INTERACTION BETWEEN STUDENT PERSONALITY AND TEACHER STYLE

<u>Personality Measures</u>	<u>Learning Outcome Measures</u>
Need-Achievement	SIAT No. 1)
*Need-Cognition	SIAT No. 2)
Berlak School Work Habits	} general measures of analytic skills in political controversy
Submissiveness	SIAT No. 3)
Guilford-Zimmerman Survey	SIAT No. 4)
*Ascendance	UNIT TESTS (homemade tests by teachers developed after each experimental unit)
*Sociability	
Emotional Stability	HEADLINES H)
Objectivity	HEADLINES G)
*Friendliness	} measures of interest in social and political issues
Thoughtfulness	
*F-Scale	
Sociometric Status	
*S-D Dichotomization	
*Tabular results presented in this chapter.	

TABLE 13.5

SUMMARY OF TWO-WAY ANALYSIS OF VARIANCE RESULTS SHOWING EFFECTS
OF RECITATION AND SOCRATIC TEACHER STYLES AND STUDENT NEED-COGNITION
ON FIVE LEARNING OUTCOMES WITH CORRELANCE ADJUSTMENT
FOR I.Q. AND POSITIVE SCORES

Means for SIAT No. 1

	Need-Cognition			Style Means	Effect	df.	F ^a
	Low	Med	H1				
Socratic Group	11.44	12.00	14.23	12.56	Style	/	
Recitation Group	11.68	10.95	14.01	12.21	Personality	2/89	6.67
Total	11.56	11.48	14.12		Style x Pers.	/	b

Means for SIAT No. 2

	Need-Cognition			Style Means	Effect	df.	F
	Low	Med	H1				
Socratic Group	41.69	46.30	52.52	46.57	Style	/	
Recitation Group	46.02	48.15	48.56	47.61	Personality	2/88	11.50
Total					Style x Pers.	2/88	4.81

Means for UNIT Tests

	Need-Cognition			Style Means	Effect	df.	F
	Low	Med	H1				
Socratic Group	151.69	154.00	163.80	156.72	Style	/	
Recitation Group	147.75	155.19	187.54	163.49	Personality	2/89	6.62
Total	149.72	154.92	175.67		Style x Pers.	2/89	2.07

Means for SIAT No. 3

	Need-Cognition			Style Means	Effect	df.	F
	Low	Med	H1				
Socratic Group	20.92	32.30	30.33	31.85	Style	/	
Recitation Group	26.15	28.41	35.26	29.94	Personality	2/88	6.33
Total	26.53	30.31	35.80		Style x Pers.	/	

Means for SIAT No. 4

	Need-Cognition			Style Means	Effect	df.	F
	Low	Med	H1				
Socratic Group	17.09	19.52	24.34	20.32	Style	/	
Recitation Group	20.91	17.46	19.55	18.97	Personality	/	
Total	18.94	18.52	21.54		Style x Pers.	2/85	1.62

a) Significance levels for 2 and 80 degrees of freedom:
.05 = 3.11 .01 = 6.96

b) Cells in which ratio was less than one were not computed.

procedures. This finding is reversed for the socratic interview measure (SIAT No. 4). Our experience in the program indicates that this finding is reasonable in view of the fact that what was measured on the unit tests bore a very immediate relationship to what had been taught in the experimental curriculum. And, for the student who is interested in "doing well" on the tests, the recitation procedure had a distinct advantage. It offered more opportunity to ask conventional questions of clarification and interpretation of what had been happening in class. The high need-cognition student in the socratic classes, in fact, must have been somewhat frustrated in his attempt to fathom what was going on, and in figuring out how to "do well on tests." The evidence indicates, however, that for the high need-cognition student this constant effort to "con" the curriculum paid off when he faced a socratic confrontation, for he did substantially better on the socratic interview measure. Moreover, while one would expect the students trained in the socratic procedure to do substantially better in the socratic interview test, this was not the case for all students -- only for a particular type of student.

Table 13.6 indicates that High F or authority-oriented students tend to score lower on tests than Low F students. This is true for all measures except the Unit Tests. Here, as with need-cognition, the general trend is reversed. This suggests again that when the test measures the results of immediate training or coaching, the more authority-oriented student is able to do at least as well, if not better, than the more independent-minded student. However, when the student faces an entirely fresh type of testing situation, the less authority-oriented student tends to do better. One additional wrinkle is added when we look at the results of the socratic interview test. While students with low and medium scores on the F-Scale tend to do better than the High F student, which is consistent with the general trend, there is a significant interaction between Low F and socratic training. That is, apparently the Low F student benefits more from socratic training than the High F student as measured by the socratic interview, while this is not the case for students trained by recitation procedures. In general, then, we might say that Low F or non-authority-oriented students tend to benefit more from training in the analysis of public issues than High F students do when the test demands some degree of transfer (when the test is not based directly on prior coaching.) Moreover, Low F students trained by a socratic procedure have a special advantage in the socratic test situation. It is particularly noteworthy that High F students trained by socratic procedures do so poorly in the socratic interview test. This is generally consistent with the findings of Stern and Cope, who suggested that the authoritarian student tends to respond better to a more permissive classroom environment.

Table 13.7 is difficult to interpret because of the absence of consistent trends. While there are three significant findings, they do not present any coherent picture. The one rather striking finding is the difference between socratic-trained and recitation-trained low dichotomizers on the socratic interview. As with the F-Scale there is a significant interaction between teaching style and student personality in which, in this case, recitation teaching does very poorly by the low dichotomizer. Our interpretation is that

TABLE 13.6

SUMMARY OF TWO-WAY ANALYSIS OF VARIANCE RESULTS SHOWING EFFECTS OF RECITATION AND SOCRATIC TEACHER STYLES AND F-SCALE ON FIVE LEARNING OUTCOMES WITH COVARIANCE ADJUSTMENT FOR I.Q. AND PRE-TEST SCORE

Means for SIAT No. 1						
	F-Scale			Style	Effect	df. F
	Low	Med	Hi	Means		
Socratic Group	12.99	12.87	11.16	12.34	Style	/
Recitation Group	13.50	12.26	11.28	12.35	Personality	2/83 2.95
Total	13.25	12.57	11.22		Style x Pers.	/
Means for SIAT No. 2						
	F-Scale			Style	Effect	df. F
	Low	Med	Hi	Means		
Socratic Group	45.13	48.18	44.85	45.06	Style	/
Recitation Group	48.90	47.03	45.04	46.99	Personality	2/82 1.60
Total	47.01	47.61	44.95		Style x Pers.	2/82 1.30
Means for UNIT Tests						
	F-Scale			Style	Effect	df. F
	Low	Med	Hi	Means		
Socratic Group	140.13	159.98	169.30	155.47	Style	/
Recitation Group	156.18	162.76	159.45	159.47	Personality	2/83 2.07
Total	148.15	161.37	164.38		Style x Pers.	2/83 1.13
Means for SIAT No. 3						
	F-Scale			Style	Effect	df. F
	Low	Med	Hi	Means		
Socratic Group	32.29	32.86	26.08	30.41	Style	/
Recitation Group	31.32	31.09	27.36	29.92	Personality	2/82 3.34
Total	31.80	31.98	26.72		Style x Pers.	/
Means for SIAT No. 4						
	F-Scale			Style	Effect	df. F
	Low	Med	Hi	Means		
Socratic Group	25.60	20.22	14.82	20.21	Style	/
Recitation Group	17.29	24.75	18.37	20.40	Personality	2/79 2.71
Total	21.30	22.56	16.66		Style x Pers.	2/79 3.35

TABLE 13.7

SUMMARY OF TWO-WAY ANALYSIS OF VARIANCE RESULTS SHOWING EFFECTS
OF RECITATION AND SOCRATIC TEACHER STYLES AND DICHOTOMIZATION
ON FIVE LEARNING OUTCOMES WITH COVARIANCE ADJUSTMENT
FOR I.Q. AND PRE-TEST SCORE

Means for SIAT No. 1							
Dichotomization				Style			
	Low	Med	Hi	Means	Effect	df.	F
Socratic Group	13.10	12.31	11.78	12.40	Style	/	
Recitation Group	12.88	13.47	10.28	12.21	Personality	2/89	3.40
Total	12.99	12.89	11.03		Style x Pers.	2/89	1.23
Means for SIAT No. 2							
Dichotomization				Style			
	Low	Med	Hi	Means	Effect	df.	F
Socratic Group	45.87	44.29	47.18	46.11	Style	/	
Recitation Group	48.06	47.37	44.86	46.76	Personality	/	
Total	47.46	45.83	46.02		Style x Pers.	2/88	1.59
Means for UNIT Tests							
Dichotomization				Style			
	Low	Med	Hi	Means	Effect	df.	F
Socratic Group	154.06	156.39	170.45	160.30	Style	/	
Recitation Group	159.25	173.48	162.11	164.95	Personality	/	
Total	156.66	164.94	166.28		Style x Pers.	2/88	1.17
Means for SIAT No. 3							
Dichotomization				Style			
	Low	Med	Hi	Means	Effect	df.	F
Socratic Group	32.12	28.15	32.26	30.84	Style	/	
Recitation Group	27.57	33.10	27.14	29.27	Personality	/	
Total	29.77	30.71	29.62		Style x Pers.	2/84	3.16
Means for SIAT No. 4							
Dichotomization				Style			
	Low	Med	Hi	Means	Effect	df.	F
Socratic Group	19.71	18.45	18.30	18.82	Style	/	
Recitation Group	12.48	22.79	20.36	18.54	Personality	2/79	2.46
Total	15.97	20.69	19.37		Style x Pers.	2/79	3.22

the dichotomization measure may get at something like affective or cognitive spontaneity. It may measure a proneness toward a more intense reaction to a problem situation as well as to individual terms or words. If this is the case, low dichotomizers may well react to a dull or low affect testing situation in a more sluggish way than high dichotomizers. On the other hand, the socratic teaching procedure forces the low affect student to react and to face an adversarial situation. This, perhaps, explains why he then does so much better in the socratic testing situation.

The major finding in Table 13.8 is a restatement of the initial finding in the correlation matrix regarding the significant relationship between G-Z Ascendancy and doing well on the socratic interview. Two additional findings are suggested by these data. The relationship holds, though not so striking, for the SIAT No. 3, which is also an interview measure, but with much less interpersonal stress. And second, low ascendant students with socratic training, while they still do poorly on the socratic interview, are not quite so handicapped as their counterparts with recitation training. We might also recall, at this point, that the correlation between SIAT No. 3 and No. 4 is .38. Earlier we speculated about whether this correlation was entirely due to a common relationship to a personality factor. (No. 3 correlates .30 with G-Z Ascendancy and No. 4 correlates .29 with G-Z Ascendancy.) While the correlation is similar, the pattern of the relationship suggests that training does make a difference (although not a statistically significant difference with this small sample).

Table 13.9 indicates that there is a clear and consistent trend toward higher sociometric status leading to high achievement. It is important to emphasize that this relationship stands up despite the fact that covariance adjustments for pre-test scores (where available) and I.Q. score have been made. The one interaction finding worth mentioning is that students with low sociometric status tend to score substantially lower on the unit test as a result of socratic teaching than the same type of student who has been exposed to recitation teaching. This finding, however, is not consistent across measures.

As indicated in Table 13.10 G-Z Sociability yields essentially the same results as G-Z Ascendancy. Low sociability, as measured by the G-Z scale is a definite liability in the student's attempt to do well on the socratic interview measure. Although there is also the same tendency toward an interactive effect (low sociability recitations doing worse than low sociability socratics), it is not nearly so pronounced.

Table 13.11 presents some rather interesting and baffling findings. At first glance it appears that there is a consistent relationship between scoring high on the G-Z Friendly scale and doing well on the learning outcome measures, although this relationship is significant in only one instance (SIAT No. 1). Upon closer inspection, we see, however, that this finding is sharply reversed in the socratic interview measure almost to the point of statistical significance. The more interesting finding that emerges is the

TABLE 13.8

SUMMARY OF TWO-WAY ANALYSIS OF VARIANCE RESULTS SHOWING EFFECTS OF RECITATION AND SOCRATIC TEACHER STYLES AND G-2 ASCENDANCE ON FOUR LEARNING OUTCOMES WITH COVARIANCE ADJUSTMENT FOR I.Q. AND PRE-TEST SCORE

Means for SIAT No. 1

G-2 Ascendancy							
	Low	Med	Hi	Style Means	Effect	df.	F
Socratic Group	13.70	11.53	12.15	12.46	Style	/	
Recitation Group	12.13	13.27	11.92	12.44	Personality	/	
Total	12.91	12.40	12.03		Style x Pers.	2/89	1.86

Means for UNIT Tests

G-2 Ascendancy							
	Low	Med	Hi	Style Means	Effect	df.	F
Socratic Group	154.50	149.14	172.14	158.42	Style	/	
Recitation Group	157.37	157.66	164.89	153.97	Personality	2/89	1.23
Total	186.68	158.40	168.51		Style x Pers.	2/89	1.25

Means for SIAT No. 3

G-2 Ascendancy							
	Low	Med	Hi	Style Means	Effect	df.	F
Socratic Group	29.05	32.25	32.95	31.42	Style	/	
Recitation Group	28.31	30.40	32.44	32.68	Personality	/	1.27
Total	28.66	31.26	32.68		Style x Pers.	/	

Means for SIAT No. 4

G-2 Ascendancy							
	Low	Med	Hi	Style Means	Effect	df.	F
Socratic Group	15.27	20.87	23.51	20.02	Style	/	
Recitation Group	12.29	21.79	22.70	18.93	Personality	2/79	4.78
Total	15.18	21.25	23.09		Style x Pers.	/	

TABLE 13.9

SUMMARY OF TWO-WAY ANALYSIS OF VARIANCE RESULTS SHOWING EFFECTS
OF RECITATION AND SOCRATIC TEACHER STYLES AND SOCIOMETRIC STATUS
ON FIVE LEARNING OUTCOMES WITH COVARIANCE ADJUSTMENT
FOR I.Q. AND PRE-TEST SCORE

Means for SIAT No. 1

	Sociometric Status			Style	Effect	df.	F
	Low	Med	H1	Means			
Socratic Group	10.55	13.46	14.58	12.87	Style	/	
Recitation Group	12.17	12.44	14.69	13.10	Personality	2/68	7.64
Total	11.20	13.05	14.62		Style x Pers.	2/68	1.09

Means for SIAT No. 2

	Sociometric Status			Style	Effect	df.	F
	Low	Med	H1	Means			
Socratic Group	42.29	47.58	47.84	47.24	Style	/	
Recitation Group	46.80	47.80	48.73	47.78	Personality	/	
Total	46.50	47.67	48.20		Style x Pers.	/	

Means for UNIT Tests

	Sociometric Status			Style	Effect	df.	F
	Low	Med	H1	Means			
Socratic Group	135.02	167.64	176.87	160.04	Style	/	
Recitation Group	160.16	154.98	186.49	167.21	Personality	2/68	7.93
Total	145.43	162.58	180.71		Style x Pers.	2/68	2.14

Means for SIAT No. 3

	Sociometric Status			Style	Effect	df.	F
	Low	Med	H1	Means			
Socratic Group	28.21	31.46	36.94	32.20	Style	1/57	1.17
Recitation Group	24.66	32.90	32.12	29.89	Personality	2/57	5.15
Total	26.79	32.03	35.02		Style x Pers.	/	

Means for SIAT No. 4

	Sociometric Status			Style	Effect	df.	F
	Low	Med	H1	Means			
Socratic Group	18.01	18.09	24.71	20.27	Style	/	
Recitation Group	20.59	18.30	23.88	20.92	Personality	2/64	1.93
Total	19.08	18.18	24.36		Style x Pers.	/	

TABLE 13.10

SUMMARY OF TWO-WAY ANALYSIS OF VARIANCE RESULTS SHOWING EFFECTS
OF RECITATION AND SOCRATIC TEACHER STYLES AND G-Z SOCIABILITY ON FIVE
LEARNING OUTCOMES WITH COVARIANCE ADJUSTMENT
FOR I.Q. AND PRE-TEST SCORE

Means for SIAT No. 1

	G-Z Sociability			Style	Effect	df.	F
	Low	Med	Hi	Means			
Socratic Group	12.21	12.50	12.27	12.35	Style	/	
Recitation Group	11.68	12.63	11.25	11.55	Personality	/	
Total	11.95	12.31	11.27		Style x Pers.	/	

Means for SIAT No. 2

	G-Z Sociability			Style	Effect	df.	F
	Low	Med	Hi	Means			
Socratic Group	45.03	48.35	47.21	45.86	Style	/	
Recitation Group	46.44	48.34	47.97	47.50	Personality	2/91	1.62
Total	45.71	48.34	47.50		Style x Pers.	/	

Means for UNIT Tests

	G-Z Sociability			Style	Effect	df.	F
	Low	Med	Hi	Means			
Socratic Group	146.41	164.24	167.49	159.38	Style	/	
Recitation Group	156.58	165.90	163.75	162.00	Personality	2/92	1.99
Total	151.34	165.04	165.68		Style x Pers.	/	

Means for SIAT No. 3

	G-Z Sociability			Style	Effect	df.	F
	Low	Med	Hi	Means			
Socratic Group	30.25	30.06	33.55	31.29	Style	/	
Recitation Group	29.43	32.46	27.19	29.69	Personality	/	
Total	29.86	31.22	30.47		Style x Pers.	2/91	1.62

Means for SIAT No. 4

	G-Z Sociability			Style	Effect	df.	F
	Low	Med	Hi	Means			
Socratic Group	16.05	23.41	19.14	19.54	Style	/	
Recitation Group	14.66	25.10	18.68	19.48	Personality	2/38	6.05
Total	15.40	24.20	18.93		Style x Pers.	/	

TABLE 13.11

SUMMARY OF TWO-WAY ANALYSIS OF VARIANCE RESULTS SHOWING EFFECTS
OF RECITATION AND SOCRATIC TEACHER STYLES AND G-2 FRIENDLINESS
ON FIVE LEARNING OUTCOMES WITH COVARIANCE ADJUSTMENT
FOR I.Q. AND PRE-TEST SCORE

Means for SIAT No. 1

	G-2 Friendliness			Style	Effect	df.	F
	Low	Med	Hi	Means			
Socratic Group	11.40	12.62	13.25	12.10	Style	1	
Recitation Group	9.05	13.61	12.23	11.90	Personality	2/56	6.39
Total	10.52	13.19	13.53		Style x Pers.	2/56	2.39

Means for SIAT No. 2

	G-2 Friendliness			Style	Effect	df.	F
	Low	Med	Hi	Means			
Socratic Group	45.18	47.49	49.55	47.41	Style	1	
Recitation Group	45.42	50.04	49.97	48.47	Personality	2/15	2.70
Total	45.31	48.95	49.79		Style x Pers.	1	

Means for UNIT Tests

	G-2 Friendliness			Style	Effect	df.	F
	Low	Med	Hi	Means			
Socratic Group	171.34	166.51	157.15	158.34	Style	1	
Recitation Group	143.58	165.04	171.47	160.30	Personality	1	
Total	155.48	157.56	165.35		Style x Pers.	2/55	4.02

Means for SIAT No. 3

	G-2 Friendliness			Style	Effect	df.	F
	Low	Med	Hi	Means			
Socratic Group	30.95	26.34	31.77	29.55	Style	1	
Recitation Group	26.77	32.03	27.24	28.68	Personality	2/52	2.24
Total	28.44	29.95	29.05		Style x Pers.	1	

Means for SIAT No. 4

	G-2 Friendliness			Style	Effect	df.	F
	Low	Med	Hi	Means			
Socratic Group	17.25	20.10	22.74	20.03	Style	1	
Recitation Group	29.07	16.48	17.67	21.34	Personality	2/49	2.47
Total	24.56	18.01	19.80		Style x Pers.	2/49	4.75

different interaction pattern for the Unit Tests versus the socratic interview. As has been pointed out before, these two tests represent the extremes in ability to transfer the material learned: the Unit Tests are most sensitive to direct coaching; the socratic interviews the least. The results indicate that high G-Z Friendly students tend to do well on the Unit Tests if they have had recitation training and poorly if they have had socratic training. Conversely, high G-Z Friendly students tend to do well in the socratic interview if they have had socratic training; and it is the low G-Z Friendly students who do well on the socratic interview if they have had recitation training. This "reverse effect" between the Unit Tests and the socratic interview has been noted before, especially in the discussion of the relationship between High F (authority-oriented) students and teaching style. It may be caused by at least two factors: the fact that the Unit Tests require the least general or transferable knowledge in order to do well, while the socratic interview requires the most general and deepest understanding of the analytic process in order to do well. (In the socratic interview you cannot think precisely about what you are doing; it has to be more or less automatic.) Or secondly, additional factors undoubtedly come into play in a face to face encounter to distract different types of students, especially when the encounter is adversarial in nature.

Kain's Study

Before concluding this chapter we should include data from a related study carried out by Kain¹⁷ within the framework of the Project using students who participated in the experimental program. Kain was interested in whether or not abstract measures of flexibility or closure might be related to the student's ability to handle controversial public questions evaluated in terms of learning outcome measures developed by the Project. The following statement indicates her rationale for selecting measures and for exploring their relationship to controversy analysis.

On the basis of the types of measures that, in the factor analytic studies of flexibility and closure, had had significant loadings on the various flexibility and closure factors, those factors were assigned to the following major categories:

- A. Factor definitions which considered as an undesirable trait the inability to break with old habits and form new ones....
- B. Factor definitions which put a negative value on the inability to break an experimental gestalt and form another new gestalt....
- C. Factor definitions which put a positive value on the ability to form one structure, discard it, and form another....

¹⁷Helen Kain, "A Study of the Relationship between Flexibility and Closure and Controversy Analysis," Cambridge, Mass. An unpublished doctoral dissertation (in process) at the Harvard Graduate School of Education, 1962.

- D. Factor definitions which considered as a desirable trait the ability to introduce structure where there had been no structure before....
- E. Factor definitions which placed a positive value on the ability to resist distraction....
 - 1. Definitions in which the resistance to distraction required only passive persistence, "holding against" or "going ahead against" distraction....
 - 2. Definitions in which the resistance to distraction required active "organizing" or "calling up and selecting out"....

Consideration of the controversy analysis task suggests that the flexibility required for rational controversy analysis might fall into any of these factor categories. A reasonable social problem-solver should be able to break with highly practiced habits of problem-analysis and solution, if evidence about a particular problem indicates that these habits are unrealistic for that problem. He should be able to break the gestalt that is his perception of the facts of the problem situation, and form a new gestalt consistent with relevant incoming data. If he has entered the problem situation without a gestalt or perceptual structure for the facts relating to the problem situation, he should be able to form one, discard it, and form another consistent with relevant incoming data. If he begins the analysis of a problem situation that is relatively unstructured as yet, he should be able to introduce structure in the form of a plan for problem analysis and, eventually, problem solution. In analyzing controversial social problems, the problem-solver must be able to hold to this plan for problem analysis in the face of distracting incoming data, and to organize the incoming data and select out relevant data in such a way that the new mass of information is no longer distracting.

Because the kinds of flexibility necessary for rational problem analysis include every major category of flexibility factor found by the factor analytical studies, it seemed reasonable to include among the experimental measures for this study tests of flexibility and closure which have had significant loadings on factors in all these categories.¹⁸

Kain then describes the nature of the tests used in her study. The essential definitions of these tests are excerpted below.

The Anagrams Test. This measure requires that the subject select the letters in five-letter words which would begin different five-letter words, if the letters in each word were rearranged.

¹⁸Ibid., pp. 33-34.

The Anagrams score is the total number of correct words identified.

Hidden Pictures test. This measure requires that the subject find and circle faces or figures of human beings hidden in the detail of line drawings. There are two Hidden Pictures scores: a fluency score, the total number of faces or figures the subject circles, and a flexibility score, the total number of faces or figures the subject circles correctly. Inasmuch as Hidden Pictures presumably tests the ability to form tentative cognitive structures, discard them, and form others until as many as possible of the hidden pictures are found, as well as the ability to break well-established perceptual habits, it might be expected that Hidden Pictures would be heavily loaded on factors in categories A and C.

Mutilated Words test. This measure requires that the subject recognize and write out familiar English words printed in small letters, parts of which are missing. The Mutilated Words score is the total number of words correctly identified. This presumably tests the individual's ability to depart from the highly-practiced habit of reading words printed in a completely legible way, so that it might be expected to have significant loadings on factors in category A. Mutilated Words also might get at the ability to form hypotheses about what the partially legible words might be, discard those hypotheses if evidence in the form of legible letter strokes indicates they are wrong, and form new ones.

Object Naming test. This measure requires that the subject name as many objects as possible that belong to each of two classes, liquids and plants. The Object Naming Test is scored for the total number of acceptable objects named, the total number of different sub-classes of objects named, and the total number of times he shifts between different sub-classes of objects. Object Naming might be expected to test ability to introduce structure where there was little or no structure before (category D), and the ability to form one structure, discard it, and form another (category C).

Teacher Rating for Dealing with Inconsistency. This is a Project developed measure, and simply asked the teacher to rate his own students on their ability to deal with inconsistency on factual, value, legal or definitional issue of a controversy by qualifying one's position, rather than by denying the inconsistency, reversing the position, or withdrawing from the argument. The process of qualification was considered essential to intelligent controversy analysis.

Interpretation of Kain's findings. Several interesting findings may be seen in Table 1312 First, in general, there is apparently no consistent or striking relationship between general measures of flexibility and the learning outcome measures. Second, in two cases there is a low but significant positive relationship between what might be considered verbal fluency (total object naming) and SIAT No. 3 and WAGMIS. For "liquids" these relationships are .20 and .35 respectively; for "plants" they are .33 and .27. One can

TABLE 13.12*

Partial Correlations

Flexibility and Controversy Analysis Scores
With Intelligence (Verbal Abilities) Held Constant

Flexibility Measures	Controversy Analysis Measures	Correlations with Verbal Abilities Held Constant
Anagrams	SIAT No. 1	-0.01
	SIAT No. 2	0.03
	SIAT No. 3	-0.02
	WAGMIS	0.01
Hidden Faces Flexibility	SIAT No. 1	0.15
	SIAT No. 2	0.08
	SIAT No. 3	0.02
	WAGMIS	0.06
Mutilated Words	SIAT No. 1	0.21 (.05)
	SIAT No. 2	0.06
	SIAT No. 3	-0.07
	WAGMIS	0.26 (.01)
Object Naming Liquids Categories	SIAT No. 1	-0.13
	SIAT No. 2	0.09
	SIAT No. 3	0.15
	WAGMIS	0.11
Object Naming Liquids Shifts	SIAT No. 1	-0.06
	SIAT No. 2	0.15
	SIAT No. 3	0.17
	WAGMIS	0.23 (.05)
Object Naming Liquids Total	SIAT No. 1	0.15
	SIAT No. 2	0.18
	SIAT No. 3	0.20 (.05)
	WAGMIS	0.35 (.001)
Object Naming Plants Categories	SIAT No. 1	0.17
	SIAT No. 2	0.24 (.05)
	SIAT No. 3	0.21 (.05)
	WAGMIS	0.22 (.05)
Object Naming Plants Shifts	SIAT No. 1	0.10
	SIAT No. 2	0.17
	SIAT No. 3	0.24 (.05)
	WAGMIS	0.18

*Adapted from Table 5 in Kain's manuscript.

TABLE 13.2 -- continued

Flexibility Measures	Controversy Analysis Measures	Correlations with Verbal Abilities Held Constant
Object Naming Plants	SIAT No. 1	0.14
Total	SIAT No. 2	0.00
	SIAT No. 3	0.33 (.001)
	WAGMIS	0.27 (.01)
Teacher Rating for Dealing with Inconsistency	SIAT No. 1	0.42 (.001)
	SIAT No. 2	0.42 (.001)
	SIAT No. 3	0.47 (.001)
	WAGMIS	0.30 (.01)

The last set of findings on Table 13.2 has special interest because it is really the relationship between three learning outcome measures and general reasoning measures (WAGMIS) and a criterion measure (Teacher Rating for Dealing with Inconsistency). The results indicate that the experimental learning outcome measures are getting at something similar to general reasoning and intelligence as measured by the measures not used in the Project.

General Discussion

With these experimental findings it is difficult to be overcautious in drawing very general conclusions. One of three major findings, however, might bear repeating. First, and as the first study to our knowledge that systematically indicated, the degree of transfer is important. The two tests were given immediately after the teaching of a specific unit and are most closely tied to what was actually taught; SIAT Nos. 1, 2, 3, and 4 are presented in increasingly abbreviated settings and tested increasingly further from the unit of the unit and in order to reach high. Our research indicated that when tests with different gradients of transferability are included, the results become much more complex. This suggests that our research strategy was probably not optimal. Instead of dealing with group data, it would probably have been much wiser to carry out the experiment in a clinical setting, i.e., to train individual students and administer tests individually.

The second finding is that there are clearly no easy interactive effects between teacher style and student personality as there are many others. Moreover, all of the major effects are in the area of student personality, rather than teacher style. This suggests that McDermott's work on teacher style or "method" is important, but its importance must be viewed out of the context of the type of material to be taught as well as the student to receive it.

A third finding is the fact that certain student characteristics, such as "ability to deal with the outer situation," can be used to predict learning outcomes. This finding was not one of the major findings of the project, but it is interesting to note that it is the only one that is not related to the learning outcome measures.

easily explain the consistent relationship between SIAT No. 3 and verbal fluency, since No. 3 is a timed interview measure. The relationship to WAGMIS, however, is a little more puzzling. WAGMIS is, as indicated earlier, significantly related to SIAT No. 3 (.26), but it is more highly related to SIAT No. 1 (.42) and No. 2 (.34). At this point, perhaps the best interpretation one can make is that verbal fluency, as measured by the Object Naming Test, is significantly related to a general measure of reasoning ability (WAGMIS), and bears a much smaller and less stable relationship to the specific measures of reasoning ability developed by the Project. We then might assume that SIAT No. 3 is measuring not only a specialized reasoning skill, but also verbal fluency itself because of its semi-openended interview format. In general, then, we might conclude that while the flexibility measures bear consistent positive relationships to our learning outcome measures, the correlations are so low as to have negligible educational significance.

The last set of findings on Table 13.12 has special interest because it is really the relationship between three learning outcome measures, one general reasoning measure (WAGMIS) and a criterion measure (Teacher Ratings). It indicates, moreover, that the experimental learning outcome measures are getting at something besides general reasoning and intelligence as perceived by the teachers who participated in the Project.

General Conclusions

With these kaleidoscopic findings it is difficult, if not impossible, to draw very many general conclusions. Two or three major findings, however, might bear repeating. First, this is the first study to our knowledge that systematically included tests which required degrees of transfer. The Unit Tests were given immediately after the teaching of a specific unit and are most closely tied to what was actually taught; SIAT Nos. 1, 2, 3, and 4 are presented in increasingly structured settings and demand increasingly firmer grasps of the analytic model in order to score high. Our research indicates that when tests with different gradients of transferability are included, the results become much more complex. This suggests that our research strategy was probably inappropriate. Instead of dealing with group data, it would probably have been much wiser to carry out the experiment in a clinical setting, i.e., to train individual students, and administer tests individually.

The second finding is that there are clearly as many interactive effects between teacher style and student personality as there are major effects. Moreover, all of the major effects are in the area of student personality, rather than teacher style. This suggests that McKeachie's hunch is right. Teaching style or "method" is important, but its importance cannot be assessed out of the context of the type of material to be taught or the personality of the student to receive it.

A third finding is the fact that certain student characteristics, which we commonly associate with "the better student," turn out to bear a relationship to learning even when our best measures of intelligence and verbal ability are partialled out. Need-cognition and sociometric status, for

example, are both commonly associated with the "good student." Both turn out to predict who is the effective learner and quite independently of intelligence. As we suggested earlier, it is our opinion that either there is an interaction between these characteristics and intelligence, which makes it impossible to cancel out the full effects of intelligence, or existing intelligence tests are not very effective measures of the type of academic aptitude required for our program.

A fourth finding relates to the striking relationship between ascendance and sociability and the student's performance in a stress interview. We have placed much hope in the stress interview as a more natural way of assessing the student's ability to state and support a political position, because the measure has such flexibility. The student can score well regardless of his political preferences, providing he is able to identify important issues and clarify them. Our analysis indicates, however, that the measure is also sensitive to interpersonal as well as intellectual competence. This should have been more obvious to us earlier, but at that time our major concern was in developing the instrument rather than assessing the total testing situation. It is undoubtedly possible to find ways of further developing the measure, or procedures for scoring it, which screen out some of these extraneous interpersonal effects. It is also, of course, an exciting area of research. To what extent do people "win" arguments, not because of what they say, but because of the manner and tone in which they say it?